Date of acceptance Grade

Instructor

Transfer learning methods for palaeoecology: comparing local models and global models

Han Lin

Helsinki June 11, 2018

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Han Lin

Työn nimi — Arbetets titel — Title
Transfer learning methods for palaeoecology: comparing local models and global models

Oppiaine — Läroämne — Subject
Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
	June 11, 2018	89 pages

Tiivistelmä — Referat — Abstract

In order to understand the relationship between organisms and environment, and reconstruct the environment in the past, where occurrence of animal species is known from fossils and climate is unknown, we build predictive models using machine learning algorithms. Our response variable for prediction is terrestrial net primary productivity (NPP) which represents fixed energy stored in vegetation. NPP is one of the main climate determinants and previous research has shown that NPP can be robustly predicted from dental traits of plant-eating mammals. Global occurrence of large plant-eating mammals and their dental traits are used as inputs.

Since occurrence of species, their traits and climate characteristic data are not uniformly distributed over time and geographical space, models built on all available training data may generate low prediction accuracy. To achieve accurate prediction, we propose three types of local models such that training data are similar to testing data. They are baseline models, hierarchical clustering based models(HCM) and advanced hierarchical clustering based models(AHCM). Moreover, hierarchical clustering are utilised for clustering data points in HCM and AHCM in order to find training data that match testing data the most.

Considering input data are not independently distributed over geographical space and therefore model evaluation is not trivial, we also propose vertical spatial cross validation (VSCV) for evaluating performance of predictive models as well as tuning parameters of models.

In experiments, ordinary least squares regression (OLS), decision tree, random forest, rotation forest and gradient boosting regressor are utilised in both global models and local models. Root mean squared error(RMSE) and mean absolute error(MAE) indicates performance of models. In an experiment, we apply VSCV to tune parameters of all models. The baseline is the global model with OLS and Africa continent is testing continent.

Experimental results illustrate that there are no models that can perform the best on each small geographic regions. Thus, we develop a scheme to give recommendations on selecting models on different regions. We recommend to use modified hierarchical clustering based models (MHCMs) and global models on the area of Lake Turkana. We propose MHCM as a new strategy to optimize HCMs. In addition, we discover that the prediction on data points in equatorial climate zone is most reliable and prediction error on the Africa continent is equatorial symmetric. Last but not the least, we demonstrate applicability of our models with a case study of fossil data from the Turkana Basin in Africa between 0.01 and 7 millions years ago. The trend of NPP over time for fossil is that NPP firstly decreases slowly and it reaches the lowest value at around 2 to 3 Ma. Then, NPP starts increasing and tends to be stable. NPP in time period between 4 and 7 Ma is higher than in present day.

ACM Computing Classification System (CCS):
J.2 [Physical Sciences and Engineering]: Archaeology,
I.5.3 [Clustering and similarity algorithms]

Avainsanat — Nyckelord — Keywords
machine learning, spatial cross validation, predictive models, hierarchical clustering, transfer learning

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction

It is known from evolutionary theory that organisms interact with and are influenced by physical environment [Dar09]. Relationships between organisms and environment can be described quantitatively by using mathematical models utilizing physical characteristics of organisms as features [Ž16]. Climate and other characteristics of environment can be predicted from occurrence of organisms at present-day where climate and occurrence of animal species is known. Those models are applied to the past where occurrence of animal species is known from fossil and climate is unknown. This study is aimed at building accurate predictive models that would help to analyze and understand the relationships and reconstruct the climate in the past over geological times. Understanding the past helps to understand evolutionary process over the ongoing climate change [BHG+17].

In machine learning context, we typically assume that training data are independently and identically distributed(i.i.d) [H+06]. However, the real species occurrences data are not uniformly distributed over geographic space and distribution is changing over time [ŽPEF17]. We test an idea that we can build predictive models with less training data which are selected in different ways for different geographic regions such that the training data are more similar to testing data instead of building global models that use all the available data. In our study, models built on all available data are global models and models made on a part of data that are available are local models. In accordance with probably approximately correct learning framework [Val84], the generalization error decreases when the number of training data increases, which means a model performs the best when there are infinity training data. But data utilised are not identically and independently distributed(i.i.d), generalization error may increase while the number of training data increase [H+06]. Then how to find good local models is encountered as a research question. So this study propose solutions to this problem.

Our problem setting is that given occurrences of animals and their physical traits, predictive models can be built for inferring productivity of the environment. But those models can not be applied to fossil data since species are different in the past. Instead, we make models on average traits of animal communities. Traits can be measured at present and in the past. Thus, we can apply such models to the past.

Furthermore, we propose three types of local models. They are baseline models, hierarchical clustering based models and advanced hierarchical clustering based models. The baseline is the global model with ordinary least squares regression. In local models settings, data consists of two groups and they are the testing data and rest data that would be selected as training data. Considered the rest data as a set, a group of training data is a subset of the rest data, and the criteria for selecting training data from the rest data is similarity compared to testing data.

Moreover, evaluation of such models is challenging since species occurrence data are not uniformly distributed over geographic space. Cross validation(CV) has been widely used for evaluating performance of predictive models and overcoming overfitting assuming input data are i.i.d [JWHT14]. However, species occurrences data are non-independently distributed over geographical space and they are spatially autocorrelated(SAC) so regular cross validation can mislead to overfitting since data are closely related with each other when they are close in geographical space [PPNH17] [LRPB13]. In other words, data points from nearby ended up in the training and testing pool, it would be almost as if a copy of some training data points is added to the testing data. Thus we modify CV and propose vertical spatial cross validation(VSCV) for assessing performance of regression models and we test spatial leave one out cross validation(SLOO) as shown in the paper [LRPM$^+$14]. The idea of designing VSCV is similar to SLOO. A group of data that are adjacent to test data are discarded since those data are likely to be correlated to test data.

In experiments, all models are evaluated by root mean squared error(RMSE) and mean absolute error(MAE). In order to understand contribution of our proposed local models to improving prediction accuracy on unseen testing data, we conduct the first experiment that parameters of those models are not tuned. In addition, to discover the optimal model, we apply VSCV to tune parameters of all models. Furthermore, prediction result of global models and local models are compared and discussed based on prediction error on testing data. Finally, the optimal model are applied to predict climate on fossil data in Turkana Basin as a case study. Following section shows related work of this thesis.

1.1 Related work

With development of statistical modeling, abundant research works that apply computational techniques in understanding climate in several million years ago by using all kinds of data have completed successfully. The related work of my thesis consists of mainly 4 aspects: Firstly, types of machine learning algorithms that are mainly selected in building accurate and robust predictive models to reconstruct environment or climate in the past; Secondly, techniques of discovering similar data points that are closely matched fossil data; Thirdly, useful methods for solving spatial autocorrelation that can result in overfitting of predictive models; Fourthly, other advanced techniques like transfer learning and what types of transfer learning algorithms are available to apply in this setting.

For the first aspect, regression models like ordinary least square linear regression(OLS) are commonly used in all kinds of papers for building predictive models for predicting precipitation, temperature or other climate characteristics. In the paper [EPL+10a], they applied linear regression and regression trees on an animals occurrences dataset of World Wildlife. The dataset provided distribution of animals occurrences on continuous ecoregions. The response variables were climate characteristics data of WorldClim. Moreover, they mapped animal occurrences data on squared cells(nearly 55km $\times$ 55km) on the world map. Features of input data utilised includes mean tooth crown height, mean body mass and diet. In their paper, R square was utilised for measuring performance of models and correlations between input features and response variables. They discovered that regression trees could be an optimal choice for modeling non-linear relationship and they found that the correlation of precipitations and mean tooth crown height with diet was the strongest. However, they just skipped the problem of spatial autocorrelation. Therefore their models can overfit. Then in their next work as in paper [EPL+10b], they directly applied the models created on modern data to fossil data shared same features. In addition, in the paper [LPE+12], ordinary least square linear regression was used in predicting net primary productivity, temperature and precipitation with two input features of hypsodonty(HYP) and longitudinal lophs count(LOP).

Furthermore, in the paper [FŽK+16], instead of building models on all data that were available, they built regression models on data points in a small area within 25 degrees of the equator on Africa continent. Moreover, like in paper Liping L et

al., two dental features: HYP and LOP were utilised as the input features space. A non-linear regression model was applied for predicting precipitation and PCA regression was selected for predicting temperature. Since HYP and LOP were linearly related closely, PCA was aimed at eliminating their relation. In addition, KNN was also applied for predicting temperature and precipitation, and they compared prediction results of fossil data. In their view, KNN and regression models had equal performance. Moreover, in the paper [FŽK+16], they extended the Functional Crown Types developed by [JHF96] to be 7 teeth features and built models using least angle regression on data points in small national parks in Kenya for predicting characteristics of climate. Their result proved that models built on input features of extended scheme of the Functional Crown Types could estimate the climate of those parks precisely. Therefore, most commonly selected models are linear regression like OLS and regression models like least angle regression that can select input features randomly can be also a good choice.

For the second aspect, clustering and PCA are commonly used in discovering prototype of data points. For example, in the paper [HAB15], hierarchical clustering was utilised to cluster plant-eating animals to 5 types of species using features like body mass and diet. Then combining those new types of clustering species with distribution of weights of large plant-eating mammals, they figured out patterns to predict temperature and precipitation and their results proved that clustering could improve accuracy of prediction indeed. As mentioned in the paper [FŽK+16], PCA was utilised to reduce linearly correlation of two input features by reducing 2D dimensional input space to 1D. Therefore it could also improve performance of models. In addition, in the paper [GTFŽ17], except for clustering and PCA, data mining methods like redescription mining could be also utilised to discover association rules of dental traits of large plant-eating mammals and characteristics of environment.

For the third aspect, standard cross validation are widely used in tuning parameters of machine learning models and eliminating overfitting with the assumption that data points are independent. However, in this settings, data points are spatially autocorrelated and in this project, there are also parameters of models need to be tuned. Meanwhile, if this problem is skipped, cross validation can result in overfitting. Like in papers [LPE+12] [HDFMB+07], they solved this problem by resampling some small parts of data points and the distance of those groups must be larger than a value like 5000kms and that value was a range for only OLS. Finally,

they built predictive models on those resampling data points.

In addition, this problem can also be solved by clustering based spatial cross validation as shown in the paper [RK], they proposed a spatial cross validation method that added a clustering method before standard cross validation to solve this problem. In their project, two datasets collected in 2007 in the growing season in two sites of Köthen. Their research unit for each data point was a 10×10 squared meter grid cell and each data point had attributes longitude and latitude to determine its location. They predicted yield based on six features like vegetation and fertilizer. They utilised k means to cluster data points in each datasets by using their longitude and latitude attributes and they set the value k to be 50. Then they gave the output of 50 means clustering to the input of standard cross validation. Finally they compared the prediction result of their spatial cross validation method and standard cross validation. They proved that models with standard cross validation were overfitting indeed. In my view, the advantage of this clustering based spatial cross validation is that there are no data points dropped. But the value of k is an important parameter to determine whether models will be overfitting. For example, in the extreme case, if k is equal to total number of data points, this method is literally the standard cross validation.

Furthermore, there is another method available as illustrated in papers [LRPB13] [PPNH17] [LRPM$^+$14], and it is spatial leave-one-out cross validation(SLOO). In the process of SLOO, a group of data points that were close to a data point were discarded and models were trained on data points that are not discarded. Literally, SLOO is like a special case of our proposed vertical spatial cross validation(VSCV) method as described in section 3.1. The differences are that firstly they used spacial distance measured by Euclidean distance and we use geographical distance using latitude and longitude of data points; secondly, they calculated pairwise distance of data points but we calculate distances of data points to boundaries of a fold.

For the fourth aspect, in this thesis, We apply models trained in the modern day data directly to fossil data since both present day data and fossil data share the same features with the assumption that joint probability distribution of dental features and NPP on present day data and fossil data are the same. But in many cases, joint probability distribution of input features and response variables on data points

from source domain and target domain can be different significantly. More importantly, if labels of data points in target domain are unknown like the situation in this thesis, it is a more challenging topic which is unsupervised transductive transfer learning [PY10]. In the paper [ANC07], they converted standard maximum entropy classifier and support vector machine to be transductive versions respectively. For maximum entropy classifier, they defined a transform function that made the response variable of source domain and a sample of target domain be the same scale. Then they built predictive models on data points in the source domain again. Likewise, for support vector machine, they used a sample of data points in the target domain with all data points in the source domain in the process of training models and they did this step in an iterative way. Datasets utilised were text data and they identified whether a word in a text was a protein name. Finally, they compared prediction results of transductive support vector machine and transductive maximum entropy classifier. They discovered that if there were only a few labels of data points in target source, it can improve performance of models significantly, and transductive support machine was better compared to transductive maximum entropy classifier in their setting.

There is another method for transductive transfer learning which is feature representation transfer [PY10]. As explained in the paper [BMP06], they proposed a structural correspondence learning algorithm. The most important step in the algorithm was that they found some pivot features that occurred frequently in data points in the source domain and the target domain. Then they combined a weight matrix to the vector of pivot features and data points with pivot features with weights were input to classifiers. Moreover, as described in the papers [BDP07] [BCK$^+$08], they propose another feature representation transfer method that is to map input feature spaces of source domain and target domain to higher dimension. Then they send this new input features space to classifiers.

Finally, The structure of this thesis is as follows: Section 2 describes algorithms of building proposed models. Section 3 describes algorithms of proposed model evaluation procedures. Section 4 illustrates datasets utilised and experiment setup for both proposed models and model evaluation methods. Section 5 illustrates results analysis. Section 6 is a case study for fossil data, and the optimal model is applied on fossil data for predicting characteristics of environment in ancient time. Finally, the last section is conclusion.

2 Proposed models

In this section, we propose 3 types of local models. Local models are models that are built on data selected from training data pool as shown in figure 1.

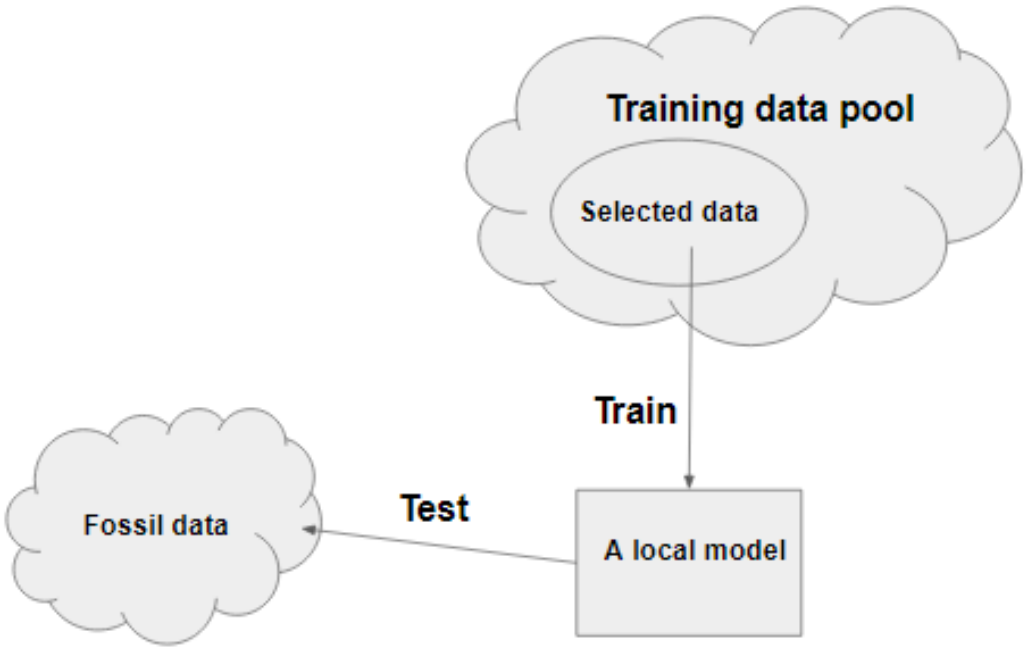

Figure 1: This figure shows the definition of local models

2.1 Predictive modeling setting

Our units of analysis instances are areas of land, such as a national park as a grid cell. Input features describe characteristics of animals occurring in those areas. The target for prediction is climate of that area, measured as productivity, rainfall and temperature variable. Assumed that we have data of some part of the modern world where animal occurrences, their characteristics as well as climate variables of those areas are known, our goal is to build predictive models that could be applied to fossil data from regions that are not the same part of the world. We will refer

to testing data as "fossil data". Besides, latitude and longitude of a data point describes location of that data point.

2.2 Local Models

In machine learning settings, input data consists of two parts: testing data and potential training data. We propose local models built on data that are selected from those potential training data. The selection criteria is based on similarity compared to fossil data, which means only data that matches fossil data closely are selected. Since data distribution spatially is not uniform. We expect that predictive accuracy would potentially be improved by selecting less training data which match fossil data more closely. Thus in this section, we propose three types of local models.

2.2.1 Baseline Models

In this section, we propose baseline models and modified baseline models. For baseline models, we manually select data points with same latitude as fossil data from training data pool. For modified baseline models, we select data points with same latitude value in both the southern hemisphere and the northern hemisphere.

To a reasonable approximation, regions with same latitude can be expected to have similar climate and environment, we expect to train the data that are located in the same level of latitude as fossil data. Firstly, two horizontal latitude boundaries can be obtained for fossil data. The top latitude boundary is the largest latitude for fossil data and the bottom latitude boundary is the smallest latitude for the fossil data. Secondly, training data have the same two boundaries as fossil data. Thirdly, a baseline model can be built on the training data by using a regressor. More precisely, a baseline model is made on a part of training data which are located in a region within two boundaries of fossil data. However, in some situations that the number of fossil data can be much larger than the number of selected training data obtained in the second step, this baseline models is not adequate since the number of the training data is too small. Thus, modified baseline models are created for improving this baseline models.

Modified baseline models(MBM) are also based on the approximation that regions with the same latitude value in both the southern hemisphere and the northern hemisphere have similar environment. This kind of baseline models are similar to

the baseline models. The first step is the same as baseline models. In the second step, training data consists of two groups. one group has the same two latitude boundaries as the testing data. The other group has two boundaries with latitude that are symmetric value of boundaries of the testing data where equator is a symmetry axis. In the last step, a MBM is built on training data obtained in the second step. We expect that by adding more training data in the second step, the accuracy of this type of baseline model can be improved.

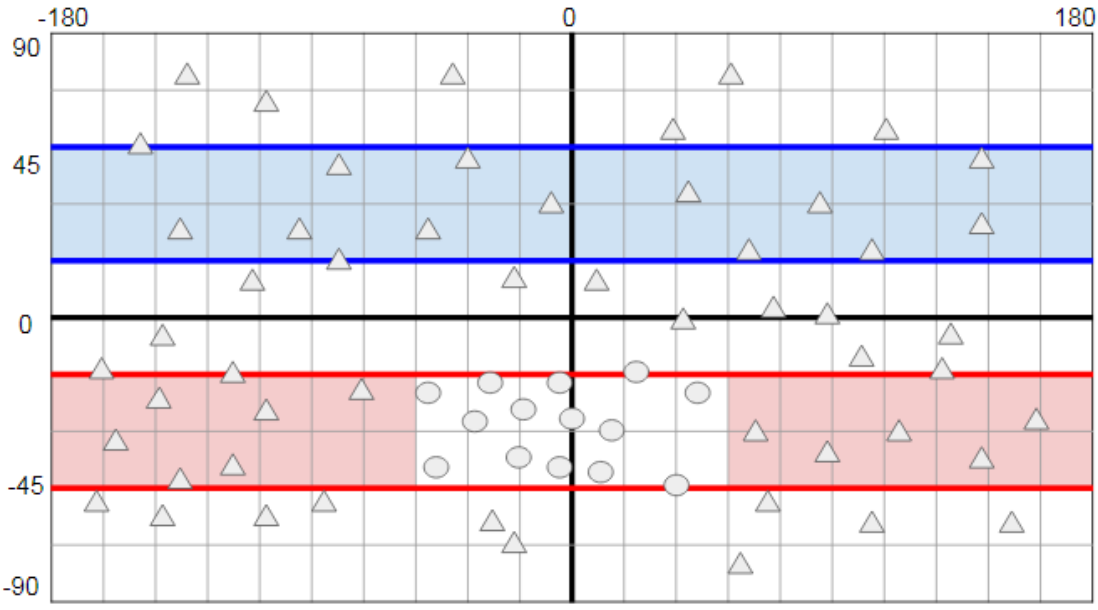

Figure 2: This figure illustrates two ways to select training data for two types of baseline models

Figure 2 describes the process of building baseline models. Circles and triangles represent input data points. Circles are fossil data and triangles are training data pool. Horizontal lines represent latitudes and vertical lines represent longitude. For type 1 baseline models, boundaries for selecting training data are those two red thick horizontal lines. Models built on triangles that are in red area between two red lines are baseline model. For modified baseline models, training data selected are triangles that are in blue area and red area.

2.2.2 Hierarchical clustering based models

In section 2.2.1, training data are manually selected data that are located in regions where climate and environment is estimated to be similar as regions where fossil data are located. Moreover, selected training data are estimated to be similar to fossil data by us. Actually, similarity between two data points can be measured by euclidean distance. Thus distance based clustering method can be utilised for finding groups of data that are similar to fossil data. Hierarchical clustering can describe how clusters are hierarchically related to each other. Thus a sequence of clusters illustrating a rank of similarity to fossil data can be obtained. So we use hierarchical clustering to automatically select data that match fossil data closely. In this section we propose hierarchical clustering based models.

Building a hierarchical clustering based model(HCM) consists of five steps. Firstly, clustering input data, including both data points in training data pool and fossil data, based on selected features to several clusters, for example k clusters. The value of k is smaller than the total number of data. Secondly, a collection of unique cluster names for testing data can be obtained, for example $S = \{x_1, x_2, ..., x_n\}, n \leq k$. Elements in the collection S are cluster names. Thirdly, started from the first element of set S, the cluster x_1 of fossil data are chosen as the testing data in the first loop. According to the result in the first step, a sequence of cluster names based on similarity compared to the x_1 cluster can be obtained, for instance $R = \{y_1, y_2, ..., y_k\}$. Fourthly, the first m clusters in training data pool are selected for building a predictive model. If we mark training data as a set T, $T = \{cluster_{y_1}, cluster_{y_2}, ..., cluster_{y_m}\}$ and $m < k$. In the fifth step, repeated the third step to the fourth step, cluster in fossil data is changed from x_2 to x_n. Therefore, hierarchical clustering based models are built for all fossil data. Moreover, value of m can be selected by using cross validation.

Figure 3 gives a simple example of process of building a hierarchical based model. All kinds of shapes in the image are data points. Different shape also represent a cluster. For example, round shapes represent cluster 1 and square shapes are cluster 2. Triangular shapes are cluster 3 and cluster 4 are represented by diamond shapes. So it means that both fossil data and data in training data pool are clustered to 4 clusters in this example. Besides, fossil data only contains one cluster which is cluster 1. Assuming that the sequence of cluster names R is $\{1, 2, 3, 4\}$ for cluster 1

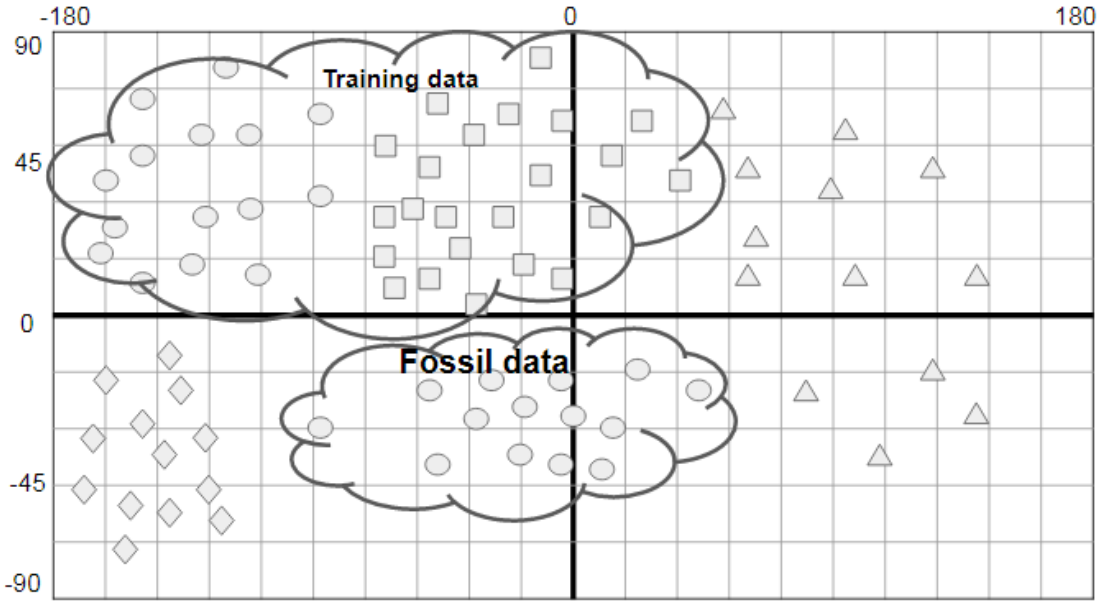

Figure 3: An example shows process of selecting training data in hierarchical clustering based models

and m parameter is chosen as 2, we select cluster 1 and cluster 2 as training data as shown in image and a hierarchical clustering model can be built on those training data. Algorithm 1 also shows process of building hierarchical clustering models.

Algorithm 1: Hierarchical clustering based models

input : Data: Fossil $\bigcup$ TrainDataPool, m

output: Hierarchical based models

Fossil [Clusters], TrainDataPool [Clusters] $\leftarrow$ `hierarchicalCluster(Data)`;
S $\leftarrow$ `Unique(`Fossil [Clusters]`)`;
for $i \leftarrow 1$ **to** `length(S)` **do**
    SubTestData $\leftarrow$ Fossil [Clusters ==S [i]];
    obtain a set R that is a sequence of cluster names for cluster S [i];
    TrainData $\leftarrow$ `ObtainTrainData(`TrainDataPool, m, R`)`;
    Model $\leftarrow$ `Regressor(`TrainData`)`;
end

2.2.3 Advanced hierarchical clustering based models

Advanced hierarchical clustering based models(AHCMs) are improved versions of HCMs. After the first step in building a HCM, it is possible that the amount of data points in a cluster in fossil data can be large and the amount of selected training data is relatively small. Thus, we expect that partitioning some large clusters in fossil data into several small parts and building models for each small part separately has potential to improve accuracy of prediction. Thus, we propose advanced hierarchical clustering based models. They are based on hierarchical clustering based models.

In the first step, clustering input data into k clusters and Select clusters of fossil data with number of data that is larger than N. We mark those selected clusters of fossil data as $S = \{clusterx_1, ..., clusterx_i\}$, where $x_i \leq k$. In addition, for the rest clusters of fossil data, HCMs can be utilised for making predictive models. In the second step, Started from cluster x_1, it is clustered into j clusters by using hierarchical clustering. In next step, started from a cluster of cluster x_1, they are concatenated to training data pool as new input data. In the fourth step, the process of hierarchical clustering based models are repeated. In this step, data that are original from the fossil data in the new input data are still a group of testing data for making predictions. Likewise, data that are original from training data pool are still a group of data that are potential training data used for building models. In last step, the second step to previous step are repeated until all clusters in fossil data have tested. Algorithm 2 shows the process from the second step to the fourth step.

Algorithm 2: Advanced hierarchical clustering based models

input : Data: fossil $\bigcup$ TrainDataPool, cluster x_i selected

output: Advanced hierarchical clustering models

```
SubTestData ← fossil [Clusters == x_i];
SubTestData [Subclusters ] ← hierarchicalCluster(SubTestData);
Unique sub-clusters ← Unique(SubTestData [Subclusters ]);
for Cluster in Unique sub-clusters do
    TestData ← SubTestData [Subclusters ==Cluster ];
    run Algorithm 1(TestData,TrainDataPool);
end
```

3 Proposed model evaluation procedures

In this section, we propose vertical spatial cross validation. Since species occurrence data are not uniformly distributed over geographic space, standard cross validation can overfit.

3.1 Vertical spatial cross validation

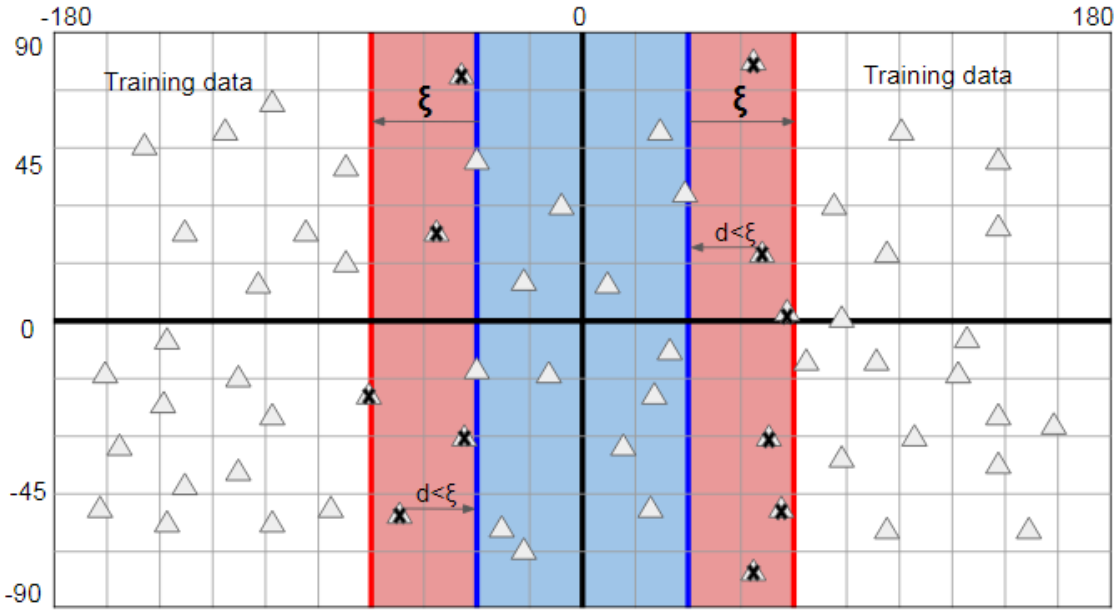

Figure 4: A summary of process in vertical spatial cross validation

This section illustrates algorithms of vertical spatial cross validation. Figure 4 and Figure 5 give examples of data distribution over geographical space. Those vertical lines represent longitudes and horizontal lines represent latitudes. Those triangles are data points. There are three steps for vertical spatial cross validation. Firstly, input data are partitioned vertically into k equal sized test folds as shown in Figure 5 and it gives an example of partitioning data into 5 folds, thus width of each fold in the image is different since data are not uniformly distributed in the geographic space. Secondly, for the a test fold as shown in Figure 4, two blue thick solid lines are boundaries for the test fold. Thus, the whole data were partitioned into three parts: the test fold, data that are on the left of the left boundary, data that are on the right of the right boundary. For data that are on the left side, those data whose

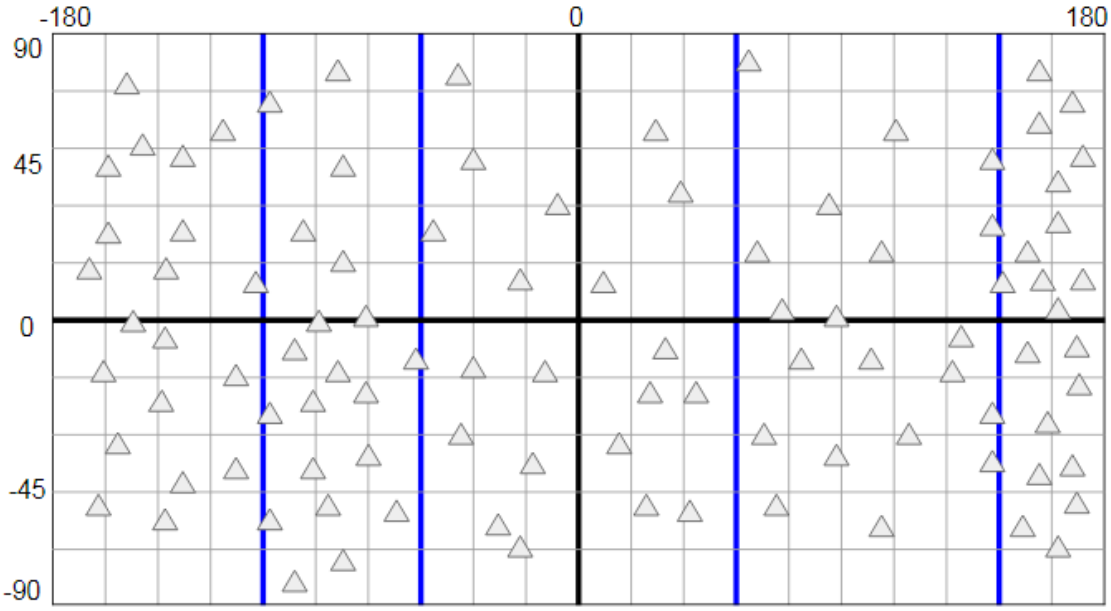

Figure 5: An example of 5 test folds

geographical distances to the left boundary are smaller than ξ are discarded; For data that are on the right side, those data whose distances to the right boundary are smaller than ξ are dropped as well. Data excluding the test fold data and data that are discarded are utilised as training data. This process is described in Figure 4. Data points that are located in the red area are discarded and there are cross signs on those data points. Thus, in the second step, started from the first fold whose left boundary has the smallest value, models can be built on their corresponding training data and prediction can be made for the first fold. Thirdly, we repeat the

second steps until all k folds are tested. Algorithm 3 illustrates the whole process.

Algorithm 3: Vertical spatial cross validation

input : Data, k, ξ

output: Error of a model

[fold 1, fold 2, ..., fold k] ← `PartitionData(Data, k)` ;

for TestFold `in` [fold *1*, fold *2*, ..., fold *k*] **do**

    leftBoundary, rightBoundary ← `GetBoundries(TestFold)` ;

    TrainDataL ← `GetTrainingDataL(Data, TestFold, leftBoundary, ξ)` ;

    TrainDataR ← `GetTrainingDataR(Data, TestFold, rightBoundary, ξ)` ;

    Model ← `Regressor(TrainDataL, TrainDataR)` ;

    prediction ← `fit(Model, TestFold)` ;

    Error ← `getError(prediction, TestFold)` ;

end

4 Experimental procedures

In this section, experimental setup for building and testing models are illustrated. Section 4.1 is a description about datasets and exact dental features utilized in experiments of this thesis. Section 4.2 illuminates steps of preprocessing datasets and technique tools utilised. Section 4.3 and section 4.4 illuminate steps and parameter settings for building models without tuning parameters and models with tuning parameters. Finally, in section 4.5, it is a description of experiment setup for VSCV that we propose, and standard cross validation as well as spatial leave one out cross validation(SLOO).

4.1 Data

This section is mainly about datasets utilised. Three datasets show animal occurrences, dental features of animals and climate variables. The fossil dataset shows mean value of dental features on locations in Turkana Basin.

In this study, three datasets are utilised for building and testing models. One of those datasets shows dental traits(taxa × traits). It describes quantitative characteristics of teeth of large plant-eating mammals. The others reveal climate for each site in

dentl traits	value
hypsodonty(HYP)	in$\{1, 2, 3\}$
longitudinal lophs count(LOP)	in$\{0, 1, 2\}$
horizodonty(HOD)	in$\{1, 2, 3\}$
acute lophs(AL)	in$\{0, 1\}$
obtuse lophs(OL)	in$\{0, 1\}$
structural fortifications of cusps(SF)	in$\{0, 1\}$
occlusal topography(OT)	in$\{0, 1\}$
coronal cementum(CM)	in$\{0, 1\}$

Table 1: dental traits [GTFŽ17]

the world(sites $\times$ bioclimate) and occurrences of taxa for each site(sites $\times$ taxa). Table 1 lists all dental traits and possible value for each type in the dental traits dataset. It is the functional dental trait scoring scheme developed in the paper [Ž16]. In dental traits dataset, it provides values of all dental traits for each taxon. The climate dataset is from the WorldClim dataset http://www.worldclim.org/. In the dataset of sites $\times$ taxa, if a taxon occurrences in a site, it is marked 1 otherwise it is 0. Thus this dataset shows taxa that appear in each site. This dataset is from the list of International Union for Conservation of Nature https://www.iucn.org/. In those two datasets, sites in Australia are excluded since dental traits of the majority herbivore in Australia are different compared dental traits of herbivore in the rest of the world [GTFŽ17]. Finally, in the case study, a fossil dataset contains mean dental traits of mammals on sits located in Turkana Basin. Moreover, those fossils are from 0.01 to 7 million years ago. We apply the optimal model which is trained on three present day datasets to fossil data for understanding the environment in the ancient time. This dataset is processed and provided by my supervisor.

In the bioclimate, animal occurrence and fossil datasets, a site represents a square grid of $50 * 50$ kilometers in the world map and it is the research unit of this work. In the sites $\times$ bioclimate dataset, there are 19 bioclimatic variables describing the climate for each site. In those variables, we use two variables: annual mean temperature(AMT) in Celsius$*10$ and annual precipitation(AP) in millimeters. As illuminated in [LPE$^+$12], NPP(net primary productivity) is the most relative variable to dental traits of plant-eating mammals since NPP measures the fixed energy stored in vegetation. NPP is calculated in the following steps: (1)$NPPt =$

$3000/(1+\exp(1.315-0.119\times AMT))$, $(2) NPPp = 3000\times(1-\exp(-0.000664\times AP))$, $(3) NPP = \min(NPPt, NPPp)$ where NPP is grams carbon in $m^{-2}year^{-1}$dry matter. In addition, fossil data are data points located in Turkana Basin in Kenya. Each data point represent a site. Fossil data also have 8 features and there are 138 data points. Furthermore, there is also a variable showing time period of the fossil.

4.2 Preprocess data

In this section, steps of aggregating animal occurrences dataset and dental traits dataset are illustrated. This aggregating process is for calculating mean dental traits on each site where there is a community of at least 3 animal species occur. Furthermore, 5 machine learning algorithms are selected and three of them are ensemble models.

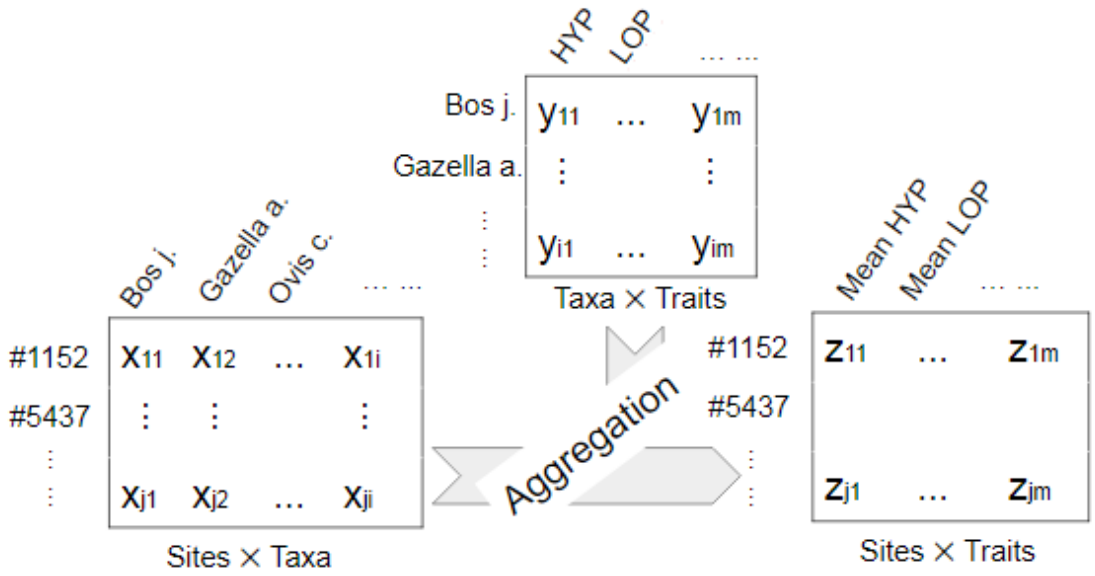

Figure 6: Data aggregation [GTFŽ17]

In the first step, data points with number of species occurred smaller than 3 are discarded, taking it account that information of dental traits in those sites are not enough for building accurate predictive models because of limited number of species [GTFŽ17]. In the next step, the dental traits dataset and occurrences of taxa dataset are aggregated to be the input dataset shown distribution of mean

value of each dental trait. In this input dataset, a mean dental feature of a data point is the average dental trait over all species occurred in a site. This process is shown in Figure 6. In this figure, all x value in sites $\times$ taxa are either 0 or 1. For any k and p, assuming $1 \leq k \leq j$ and $1 \leq p \leq m$, $z_{kp} = \frac{\sum_{n=1}^{i} x_{kn} * y_{np}}{\sum_{n=1}^{i} x_{kn}}$. However, there are several missing value for a few dental traits in the taxa $\times$ traits dataset. Those missing data are skipped in the process of aggregation. In the third step, NPP value on each site are calculated based on the formula illustrated in the section 4.1. Finally, both the input features and response value NPP for each site are ready for building models and there are 28886 number of data points. The input data reveal the mean dental traits of communities of mammals with at least 3 species and NPP reveals the environment in the present day. Thus predictive models describe the relationship between them.

In addition, five different machine learning algorithms are selected to build models and their performance on testing data are compared with each other. They are ordinary least squares regression(OLS), decision tree(CART)(DT), random forest(RaF), rotation forest(RoF), gradient boosting regressor(GBR). Regression models are selected since NPP are not discrete value like class labels. In addition, OLS and decision tree are tested because OLS are commonly used in this setting and decision tree are selected because in the paper [EPL+10a], their result shows that regression trees can have good performance. But it is easy to overfit. In addition, three ensemble models: random forest, gradient boosting regressor and rotation forest are selected. Gradient boosting regressor utilised in this thesis is decision trees as weak learners and the algorithms of the model is that every time adding a model built on residuals to previous model to minimising the least squares error [Fri01]. The advantage of this algorithm is that it contributes significantly in reducing bias. Moreover, the advantage of random forest is that it is relatively hard to overfit data points. Finally, rotation forest is selected because input features are rotated in k directions with maximum variation, which can reduce linear correlation between input features result in making accurate models [RKA]. Finally, when building clustering based models, hierarchical clustering with ward's linkage method are utilised for clustering data.

Furthermore, we choose Africa as the testing continent for both global models and local models. This is because recent Africa environment is relatively least affected by human activities and this way is expected to be similar to fossil data. The amount

	parameters
decision tree	number of depths
random forest	number of estimators
gradient boosting regressor	number of estimators, learning rate and number of depth
rotation forest	number of subsets and number of trees

Table 2: This table shows parameters for four machine learning models

of Africa data is 8235. Furthermore, the rest data points that are not located on Africa continent and Madagascar form training data pool and the amount of data points in the training data pool is 20651.

In our experiments root mean squared error(RMSE) and mean absolute error(MAE) are utilised for measuring performance of models. RMSE gives relatively high weights to large errors and large errors are undesirable in our experiment so it is used. MAE is utilised to measure the accuracy of prediction. Both RMSE and MAE are negatively-oriented scores which means the smaller the value of RMSE and MAE is, the better the model is.

In our experiment, decision tree(CART), random forest, gradient boosting regressor are libraries in sklearn in python. OLS regression is from statsmodels package. Rotation forest is tested from source codes. Parameters for four machine learning models is shown in Table 2. The python version used is 2.7.9 with 64 bit. The version of sklearn is 0.18.1 and statsmodels is 0.8.0.

4.3 Experiments for building models without tuning parameters

In this section, Parameter settings for 4 machine learning algorithms for both global models and local models are the same as shown in Table 3. In order to understand contribution of selecting training data from the training data pool for building local models in improving prediction accuracy, parameters settings of both global models and local models are the same. In addition, for both global models and local models, OLS, decision tree, random forest, gradient boosting regressor and rotation forest are tested separately.

models	parameter settings
decision tree	layers: 10
random forest	estimators: 10
gradient boosting regressor	estimators: 10,learning rate: 0.01, maximum depth is 2
rotation forest	k: 2, number of trees is 35

Table 3: parameters settings for four machine learning models before tuning parameters

4.3.1 baseline models and Modified baseline models

This section illustrates parameters settings of two types of baseline models and steps of conducting them before tuning parameters. data points on Africa continent and Madagascar are testing data and training data are selected data from the training data pool. Testing data are partitioned into ten horizontal layers and the height of each layer is not larger than 5 degrees(almost 555 kilometers) for both two types of baseline models. In addition, data points with latitude that are larger than 12.74 in the northern hemisphere are not included as testing data since there are a few data points and their distribution in the map is dispersive.

4.3.2 Hierarchical clustering based models

This section shows implementation details of hierarchical clustering based models before tuning parameters. The whole data points which consists of testing data and all data points in the training data pool are partitioned into ten clusters using hierarchical clustering. The number of clusters is 10 because the climate of Africa can be classified to 8 different zones. While number of clusters is 10, 8 clusters appear on Africa continent. Therefore, those clusters can almost correspond to 1 or 2 climate zones. Figure 7 shows distribution of different clusters in the world map and Figure 8 is the dendrogram revealing similarity of different clusters and ten leaves are the first cluster to the tenth cluster from the left corner leaf to the right corner leaf. A rank of clusters names as mentioned in section 2.2.2 is generated from this dendrogram. For example, cluster 6 is selected as testing data, a rank of cluster labels for selecting training data(without Africa data) is $\{6, 5, 9, 10, 7, 8, 3, 4, 1, 2\}$. The most similar data to testing data which are the cluster 6 in the training data.

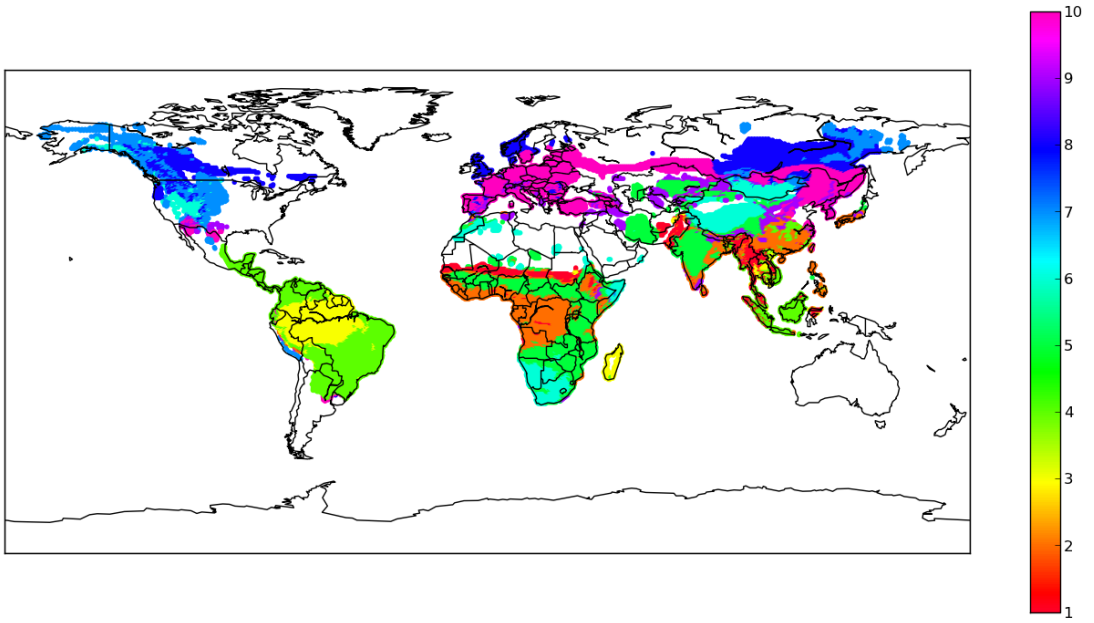

Figure 7: This figure shows distribution of 10 clusters on the world map and a color represent a cluster. The color map on the right shows corresponding cluster of a color

Furthermore, clusters in red have higher similarity than clusters in green since the cluster 6 is red and clusters in red can be merged in a larger cluster as shown in image 8. In addition, cluster 6 and cluster 5 can be merged as a cluster so the cluster 5 in training data is in the second position in the vector of the rank of cluster labels. Moreover, cluster 3 and 4 have higher rank than cluster 1 and 2 since the distance of cluster 3 and 4 is smaller than cluster 1 and 2.

For building HCMs without tuning parameters, RMSE and MAE are kept in the process of appending clusters from training data pool to training data. For example, for cluster 6 as testing data, in the first round, a HCM is built with 1 cluster data points from training data pool and prediction error is recorded. Then, in the second round, prediction error of a HCM that is built with 2 clusters data points from training data pool are kept. Then repeating this process until 10 clusters in training data pool are selected for building HCMs and when 10 clusters are selected for building models, the model is a global model as there are total 10 clusters in the training data.

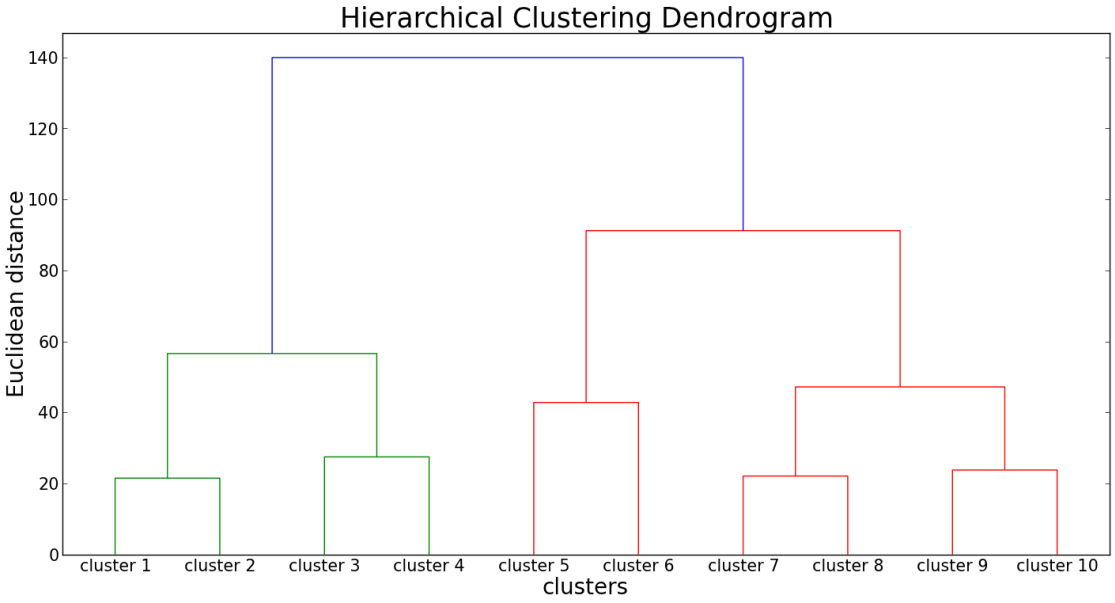

Figure 8: This figure is the dendrogram of the clustering result. This also shows how clusters can be merged into a larger one. For example, cluster 1 and cluster 2 can be merged as a cluster. In addition, this also reveal similarity of clusters and the vector R of a rank of a cluster in testing data is generated from this.

4.3.3 Modified hierarchical clustering based models

In order to improve performance of hierarchical clustering based models, an optimization strategies are utilised in making predictions. Since the number of data in a cluster in testing data can be large. Data points in a large cluster are partitioned into some small groups and parameters of models can be optimized on those groups separately.

This optimization strategy consists of several steps. Firstly, a cluster in test data is selected. Secondly, that cluster is partitioned in a horizontal way into some layers and the span of layers are almost the same. This step is the same as the first step in building the first baseline models in section 4.3.1. Thirdly, started from the first layer data of the cluster, they are testing data. As mentioned in section 2.2.2, a set R which is a rank of clusters based on similarity for a cluster can be obtained. Thus, a set $R = \{y_1, y_2, ..., y_k\}$ for this cluster can be obtained. The fourth step is the same as the step four in section 2.2.2. Started from the y_1 cluster in training data, a following cluster of the cluster in previous round is appended in next round until all training data are included in building a model. Algorithm 4 reveals the

whole process. If there are more than one clusters that need to be analyzed, this algorithm can be ran for each cluster respectively.

Algorithm 4: Local Models: Modified hierarchical clustering based models

input : Data: TestData + TrainData, m layers, cluster x_i selected

output: Modified hierarchical clustering based models

```
SubTestData ← TestData [clusters == x_i];
[layer_1, layer_2, ...layer_m] ← SubTestData;
obtain a set R that is a rank of cluster labels for cluster x_i;
for layer in [layer_1, layer_2, ...layer_m] do
    for j ← 1 to length(R) do
        TempTrain ← TrainData [clusters ==R [j]];
        TotalTrain ← Combine(TotalTrain,TempTrain);
        MHCM ← Regressor(TrainData, layer);
    end
end
```

The cluster 1 in Africa are partitioned into three horizontal layers and the span of each layer is smaller than 333 kilometers and data of cluster 1 with latitude that is smaller than 9.6 are discarded since data with latitude that is below 9.6 are distributed dispersedly and the number of those data are not large. Each layer are tested as the way in hierarchical clustering based models. Training data are selected also from the data without the Africa data. In the process of selecting training data, a rank of clusters of testing data is also generated as the order to select training data.

The cluster 2 in Africa are partitioned into 6 layers and the span of each layer is almost 555 kilometers. Data points of cluster 2 with latitude that is above 14.99 or below -14.54 are not included as testing data with the same reason mentioned above. Cluster 5 in Africa are partitioned into ten layers as the way for the cluster 2 and data of the cluster 5 with latitude that is above 12.74 are also discarded. Finally, cluster 6 are partitioned into four layers. Each layer in those clusters is tested as the same way for the cluster 1. Like in section 4.3.2, prediction error are recorded in the process of appending one cluster a time until all clusters in training data pool are selected for building models.

4.3.4 Advanced Hierarchical clustering based models

Like in modified hierarchical clustering based models, cluster 1, 2, 5 and 6 in Africa are clustered to be several sub-clusters. Cluster 1 are clustered to 3 sub-clusters. Cluster 2 are clustered to 6 sub-clusters. Cluster 5 are clustered to 10 sub-clusters. Cluster 6 are clustered to 4 sub-clusters. Each sub-clusters are tested respectively and a new set of the rank of cluster labels for a clusters is generated by combining it with the rest of data without Africa and clustering them again. In this model, there are no testing data that are discarded.

4.4 Experiments for building models with tuning parameters

Tuning parameters of models can result in optimal models for fossil data. In this thesis, for building clustering based models, number of clusters selected as training data is also a type of parameters. We use proposed VSCV to tune parameters of global models and local models. Firstly, the whole testing data are partitioned into 3 test folds like the way described in section 3.1 and number of data points in a testfold is 2745. For each test fold, in the rest data points in Africa continent, data points of which distance to any boundary of the test fold is smaller than 500km are discarded and the remaining data points are utilized as validation data. VSCV are utilized instead of standard cross validation because data points are spatially autocorrelated. Parameters of four machine learning models and number of clusters involved in training data are tuned for a test fold by minimizing RMSE of models on validation data. For example, number of depths, number of estimators of decision tree and random forest are tuned from 1 to 36 with interval value 1. Number of estimators of GBR is tuned from 1 to 30. Meanwhile, learning rate is tune from 0.0001, 0.0005, 0.001 to 0.505 with interval 0.005 and depth is tuned from 1 to 5. For rotation forest, k is tuned from 2, 4 and number of trees are tuned 5 to 20 with interval 5. Moreover, steps of building both global models and local models are the same as building models without tuning parameters but predictions are not recorded while the process of appending training data. In addition, we test three rounds until all data in Africa are acted as testing data once.

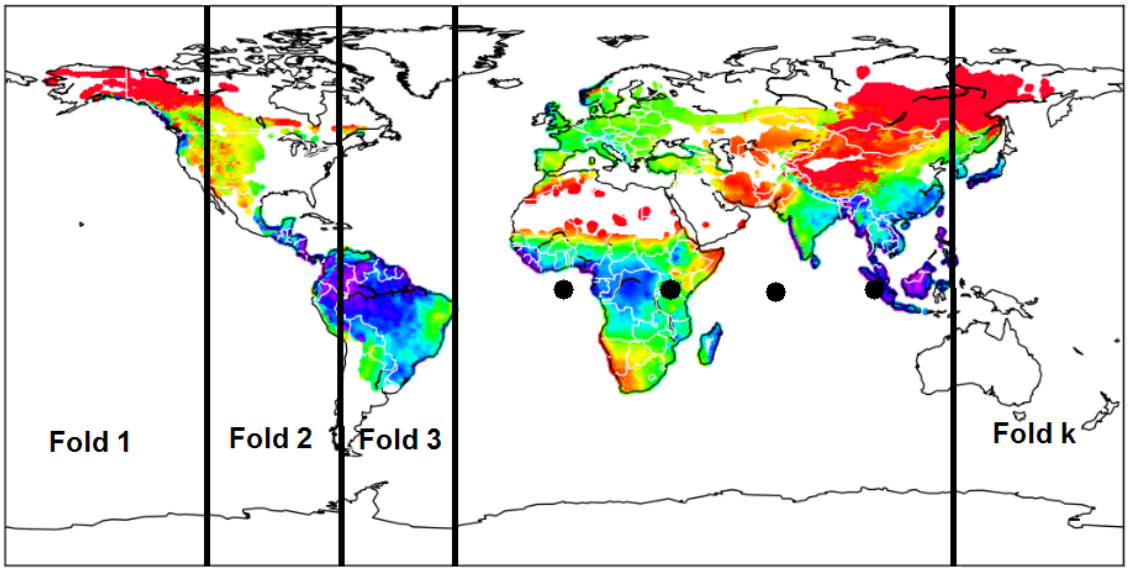

Figure 9: An example of 11 test folds on the world map

4.5 Vertical spatial cross validation and spatial leave-one-out cross validation

This section illustrates parameters settings for our proposed vertical spatial cross validation(VSCV). In addition, spatial leave-one-out cross validation(SLOO) and standard cross validation are also tested. Performance of five machine learning models using VSCV, SLOO and standard cross validation are compared individually. Moreover, parameters settings of SLOO and standard cross validation are also presented in this section.

In machine learning, cross validation is used commonly as a method to overcome overfitting and tune parameters. Thus, we train and test models as the way in cross validation. For using vertical spatial cross validation, three steps are conducted in our experiment. Firstly, input data are partitioned vertically into 11 equal sized test folds as shown in Figure 9, thus width of each fold in the world map is different since data were not equally distributed in the map. Thus there are 2626 number of data for each test fold. Secondly, for the x-th test fold as shown in Figure 4, two blue thick solid lines are boundaries for the x-th test fold. Thus, the whole data are partitioned into three parts: the test fold, data that is on the left of the left boundary, data that is on the right of the right boundary. For data that is on the

left side, those data whose distances to the left boundary are smaller than 500 kilometers are discarded; For data that is on the right side, those data whose distances to the right boundary are smaller than 500 kilometers are dropped as well. Finally, data excluding the test fold and data that are discarded are utilised as training data or validation data. Data that are located in the grey area are discarded. Thirdly, models can be built from training data for each test folds and prediction are made for each test folds. This process is marked as spatial 11 folds cross validation.

For using spatial leave-one-out cross validation, in each training and testing loop, one data point acts as testing data and some of the rest data are training data. The process contains several steps. In the first step, a data is selected as a test data. In the second step, all data points that are located in the point where is at least 500 kilometers away from the test data are training data. This process is marked as spatial leave-one-out cross validation. In addition, standard 11 fold cross validation and leave-one-out cross validation are also tested for comparison with VSCV that we propose.

Considering the running time of rotation forest is large, it is not tested in standard leave-one-out cross validation and spatial leave-one-out cross validation and the rest four models are tested. Parameters settings of four models for 11 fold cross validation are also the same as Table 4.

models	cross validation(standard and spatial)
decision tree	20 layers
random forest	normal: 10 estimators; spatial: 25 estimators
gradient boosting regressor	7 estimators,learning rate is 1.2 and maximum depth is 10
rotation forest	k is 2 and number of trees is 25

Table 4: parameters for VSCV, SLOO and standard cross validation

5 Result Analysis

This section presents prediction accuracy of global models and local models on unseen testing data. In addition, we choose global models built with OLS as baseline. Since it is the simplest and widely used model in this setting and it expects to have

good results as well. Moreover, this thesis is aimed to develop good local models that can improve prediction accuracy. Furthermore, testing data are data points in Africa continent including Madagascar and data points in Eurasia, South America and North America continents form training data pool. Thus, this section is arranged as: section 5.1 presents prediction results of global models and three types of local models that we propose. In addition, parameters of those models are not tuned. Likewise, section 5.2 shows results of them after tuning parameters. Moreover, section 5.3 is a discussion of global models and our proposed models. In addition, we develop a scheme shows optimal models on different regions of Africa. Finally, section 5.4 describes results of our proposed vertical spatial cross validation(VSCV) and we compare VSCV with standard cross validation and sptatial leave-one-out cross validation(SLOO).

5.1 Results of models before tuning parameters

This section depicts results of global models and local models before tuning parameters. In addition, we also describe results of them using OLS, Decision Tree(DT), Random Forest(RaF), Gradient Boosting Regressor(GBR) and Rotation Forest(RoF) individually. Meanwhile, parameters of DT, RaF, GBR, RoF and number of clusters selected in training data are also not tuned. Thus, we present the change of RMSE and MAE with increment of number of clusters in training data for three clustering based models.

5.1.1 global models

machine learning models	parameter settings	RMSE	MAE
OLS	–	565	430
DT	depths: 10	686	518
RaF •	estimators: 10 •	552 •	425 •
GBR	estimators: 10, learning rate: 0.01 and depth: 2	577	496
RoF	k: 2 and trees: 35	687	526

Table 5: Summary of performance of global models with five different machine learning models and their parameters settings. RaF is the best model and it is highlighted with bullets.

This section describes prediction accuracy of global models with five different machine learning algorithms and predictors with RaF before tuning parameters is the

best model.

Table 5 gives prediction accuracy of global models on testing data. Thus, the predictor with RaF is the best model and RoF is the worst. Thus, RMSE of the best model is 19.6% less than the worst model. In addition, RMSE of the baseline, which is the global model built by OLS, is 565. It is the second best model with RMSE 565. Compared to the baseline, performance of the best algorithm improves 2% even though the parameter of RaF is not tuned. RaF is better than OLS since it is a more complex model. For each prediction, it use the mean value of all weak learners. Although one weak learner can have negative prediction, the average of 10 weak learners can be good. Moreover, It is reasonable that GBR and RoF do not have well performance since their parameters are not tuned. Thus predictors with them can be underfitting. Thus, their performance can be expected to be improved after tuning parameters.

5.1.2 Baseline models

machine learning models	parameter settings	RMSE	MAE
OLS	–	7.40×10^{14}	2.62×10^{14}
DT	depths: 10	854	701
RaF •	estimators: 10 •	737 •	572 •
GBR	estimators: 10, learning rate: 0.01 and depth: 2	786	640
RoF	k: 2 and trees: 35	1262	1150

Table 6: RMSE and MAE of baseline models used OLS, DT, RaF, GBR and RoF on testing data. The best predictor is marked with bullets.

In this section, performance of baseline models are presented and RaF is proved to be the best model as well. However, performance of the best baseline models without tuning parameters is worse than the baseline. By analysing performance of the best baseline models on each layer, training data selected are not adequate can be one possible reason result in poor performance. The other reason is that selected training data are not similar to testing data in some layers.

As shown in Table 6, The best BM is built by random forest with RMSE 737. It is 1.3 times as the RMSE of the baseline. Reasons of the poor performance of baseline models can be :firstly, parameters of DT, RaF, GBR and RoF are not tuned. On other words, those models are so simple that they are underfitting. Secondly, the training data selected in baseline models are not enough or they are not close to testing data.

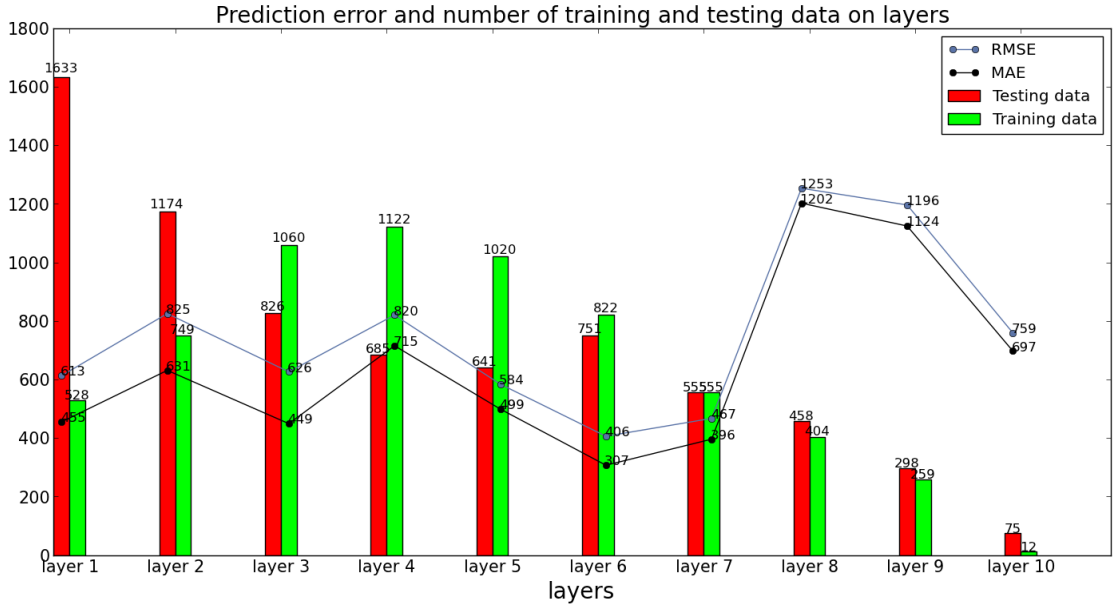

Figure 10: Prediction error of the best baseline model on testing data points and the amount of training data and testing data utilised in each layer.

As shown in Figure 10, Each layer represents a horizontal slides with span nearly 555 kilometers from the North to South of Africa. Heights of red bars and green bars represent the number of testing data and the number of training data respectively. The blue line and red line show RMSE and MAE on each layer and they are result of a baseline model utilized RaF. Moreover, in five layers: layer 1, layer 2, layer 8, layer 9 and layer 10, the number of training data is less than the amount of testing data. For example, the number of training data is 30% of testing data points in layer 1 and the RMSE of the layer is 613. Thus, One possible reason for poor performance of the optimal baseline model is that selected training data are not adequate. However, in layer 4, the number of training data is 1.6 times as the testing data and its RMSE is 820. Thus, it is possible that selected training data for testing data in layer 4 are not similar to them. This results in relatively high RMSE.

Therefore, a smarter way to select training data is required to improve performance of baseline models.

5.1.3 Modified baseline models

This section illustrates performance of the best modified baseline model. It improves prediction accuracy compared to the baseline. The advantage of modified baseline models is that less training data are involved and it shortens the running time. Then, we compare performance of the best modified baseline model on each layers with the best baseline model. Performance on all layers are improved by modified baseline models and contributions are mainly from layers in the south of Africa.

RMSE of the optimal modified baseline model is 1.4% less than the baseline. The performance of the best modified baseline model improves 24% compared to the best baseline model. The RMSE and MAE of five models are shown in Table 7. According to the result, random forest is still the best algorithm with RMSE 557 as marked in Table 7. The overall performance on the Africa continent of the best modified baseline model is the same as the best global model of which the RMSE is 552. But the advantage of local models like modified baseline models, the amount of training data utilised is less. For example, the total number of training data utilised in modified baseline model is 74% of the training data in global model but the performance of them are the same in the equal condition that parameters settings are the same as well. Thus, the running time can be shorten.

machine learning models	parameter settings	RMSE	MAE
OLS	–	1162	870
DT	depths: 10	648	514
RaF •	estimators: 10 •	557 •	412 •
GBR	estimators: 10, learning rate: 0.01 and depth: 2	740	601
RoF	k: 2 and trees: 35	1118	1007

Table 7: RMSE and MAE of modified baseline models.

Furthermore, the performance of OLS in modified baseline model is at least 10^{11}

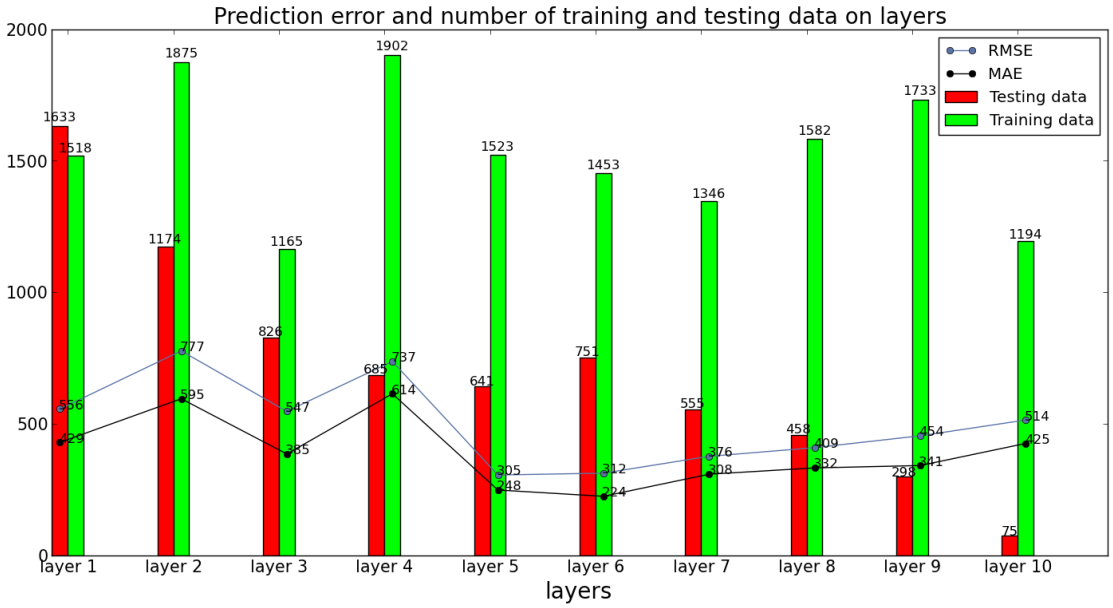

Figure 11: This figure shows RMSE and MAE of a modified baseline model built by Random forest on 10 layers on the Africa continent and the amount of training data and testing data utilised in each layer on Africa.

times better than its performance in baseline model. Thus, appending more training data that can match testing data closely can contribute a lot in improving performance of regression models. Figure 11 shows explicitly the change of the amount of training data utilised in each layer if it is compared with Figure 10. Like in Figure 10, the amount of training data and testing data are represented by the height of green bars and red bars, and RMSE and MAE shows in the figure are modified baseline model with Random forest. It is very obvious that RMSE of all layers are reduced with the increasement of training data compared to the result in Figure 10. The percentage of improvement in each layer for performance of the best modified baseline model compared to the best baseline model is shown in Table 8. Thus, the performance of the best modified model is improved on each layer and layer 8 has the most improvement which is 67%. Therefore, training data selected in baseline model are not sufficient result in poor performance of models. Furthermore, the layer 5 has the lowest RMSE in the modified baseline model. Equator is covered in layer 3. The figure shows that the best modified baseline model contributes a lot in improving performance in testing data points that are on the south of Africa. But performance of local models on testing data points in the North need more improvement.

	layer 1	layer 2	layer 3	layer 4	layer 5	layer 6	layer 7	layer 8	layer 9	layer 10
percent(%)	9	5	13	10	48	23	19	67 •	62	32

Table 8: This table shows improvement of modified baseline model compared to baseline model for each layer in testing data.

5.1.4 Hierarchical clustering based models and modified hierarchical clustering based models

Hierarchical clustering based models(HCMs), modified hierarchical clustering based models(MHCMs) and Advanced hierarchical clustering based models(AHCMs) are also local models. Their training data are selected based on the distances calculated in the process of clustering all the data points available. The shorter the distance between two different clusters, the more similar them are. Thus for each cluster in testing data, it is possible to find a group of data points from training data pool that match testing data the best.

Prediction results of MHCMs are the same as results of HCMs since the way of MHCMs to select training data is the same as HCMs and parameter settings are the same for them. Moreover, from Figure 12 to Figure 16, they represent the change of RMSE and MAE of HCMs and MHCMs by using OLS, DT, RaF, GBR and RoF over ten different types of training data. For example, in Figure 13, the height of a single red bar and a single green bar represents RMSE and MAE of a HCM or MHCM built by DT and tested on the whole testing data. Labels on x axis in figures stand for the number of clusters selected as training data to build a HCM or MHCM. For example, "2 clusters" on x axis in figures represents two clusters from the training data pool are selected for building a HCM or MHCM. Therefore, 2 bars with x axis value "10" clusters, which are the rightmost two bars in figures, represent the prediction result of global models since all clusters in the training data pool are selected for building models in that case.

In Figure 12, RMSE and MAE of a HCM or MHCM with OLS using 1 cluster in training data are not included in the figure. Since they are too large so that they are out of scale of the image. The RMSE and MAE are 4.69×10^{13} and 9.90×10^{12}.

Figure 12: This figure shows the change of prediction error of OLS over the process of appending one cluster of data from the training data pool to training data each time for building HCMs and MHCMs.

Figure 13: This figure shows the change of prediction error of DT over the process of appending one cluster of data from the training data pool to training data each time for building HCMs and MHCMs.

Figure 14: This figure shows the change of prediction error of RaF over the process of appending one cluster of data from the training data pool to training data each time for building HCMs and MHCMs.

Figure 15: This figure shows the change of prediction error of GBR over the process of appending one cluster of data from the training data pool to training data each time for building HCMs and MHCMs.

Figure 16: This figure shows the change of prediction error of RoF over the process of appending one cluster of data from the training data pool to training data each time for building HCMs and MHCMs.

The change of RMSE and MAE of Figure 12 is that when training data change from 1 cluster data points to 2 clusters data points, RMSE and MAE decrease sharply to 560 and 398. They reach the lowest value. Thus a HCM or MHCM building by OLS can have the best performance when two clusters data points are selected as training data. Then, while adding more clusters of data points to training data, the performance of the model becomes worse. RMSE and MAE reach another high value when there are four clusters in the training data. In that process, prediction error increase 36% from the lowest value. Then RMSE and MAE start decreasing with the increment of training data. Finally, when all data points in the training data pool are used for building a model, namely a global model, RMSE and MAE reach the value that are almost equal to the best value. They are 565 and 430.

In Figure 13, when there are only one clusters in the training data, prediction error is relatively high and RMSE is 790. Then it starts decreasing. When there are 3 clusters in the training data, prediction error start increasing and the highest error appears when there are 5 clusters utilised for building a model. Then prediction error begins decreasing gradually. Ultimately, when all clusters are utilised in training data, the performance of model is the best with RMSE 686. In Figure 14, it shows

the change of prediction error when building HCMs or MHCMs with Random forest with the same increment as description for Figure 13. The change of RMSE and MAE is that they increase when number of clusters in training data increasing from 1 to 3 clusters. Then they drop rapidly when number of clusters in training data increase from 3 to 4. Then they increase while number of clusters change from 4 to 5. Then they reach the lowest value when number of clusters in training data is 7 and RMSE is 550. The lowest value of prediction error of HCM or MHCM built by random forest is nearly equal to it of a global model made by random forest. When number of clusters is 10 in the training data, RMSE is 552. The change of RMSE and MAE in Figure 15 is very similar to the change of prediction error made by HCMs and MHCMs with Random forest. HCMs or MHCMs with GBR has the best performance when number of clusters in training data is 8 and RMSE is 554. The change of RMSE and MAE of HCMs and MHCMs with RoF as shown in Figure 16 is that RMSE and MAE is highest value when number of cluster is 1. Then they start to drop until number of clusters in training data is 5 and they reach the lowest value which are 579 and 486. Then they increase with augment of number of clusters in training data.

machine learning models	number of clusters in training data	RMSE	MAE
OLS	2	560	398
DT	10	686	517
RaF •	7 •	550 •	422 •
GBR	8	554	479
RoF	5	579	486

Table 9: This table shows the best prediction result and the best parameters for selecting the number of clusters in the pool of training data for five machine learning algorithms individually.

Table 9 shows the best RMSE and MAE for HCMs and MHCMs when using five different machine learning algorithms. In addition, it also illustrates the number of clusters which can lead to the best performance when using different machine learning models. The number of clusters needed for DT, RaF, GBR and RoF is larger than or equal to 5 when models reach the best performance. However, the best performance of models with OLS occurs when there are only 2 clusters in training data. In my view, it is possible that parameters of DT, RaF, GBR and RoF are

not tuned so the model itself is too simple; thus, more training data are required to train the model. As shown in the table, the best HCMs and MHCMs is built by Random forest with RMSE 550. The performance of the best HCM and MHCM is just a little bit better than the best global model with RMSE 552.

5.1.5 Advanced hierarchical clustering based models

Figure 17: This figure shows the change of prediction error of ols over the process of appending one cluster of data from the training data pool to training data each time for building AHCMs.

Figure 17 to Figure 21 reveals the change of RMSE and MAE of advanced hierarchical clustering base models with ols, DT, RaF, GBR and RoF while number of clusters in the training data increase from 1 to 10. In Figure 17, the trend of RMSE and MAE is similar as in Figure 12. The value of prediction error of AHCMs when number of clusters is smaller than 4 does not appear in the figure since they are too large, more than 10^{13}. Thus, they are out of scale of the figure. But the best performance of an AHCM with ols appears when number of clusters is 10. Namely, the best AHCM with ols is the global model with ols. The trend of RMSE and MAE of AHCMs with DT is like the trend of prediction error of HCMs with DT. But the best performance of AHCMs with DT occurs while number of clusters in

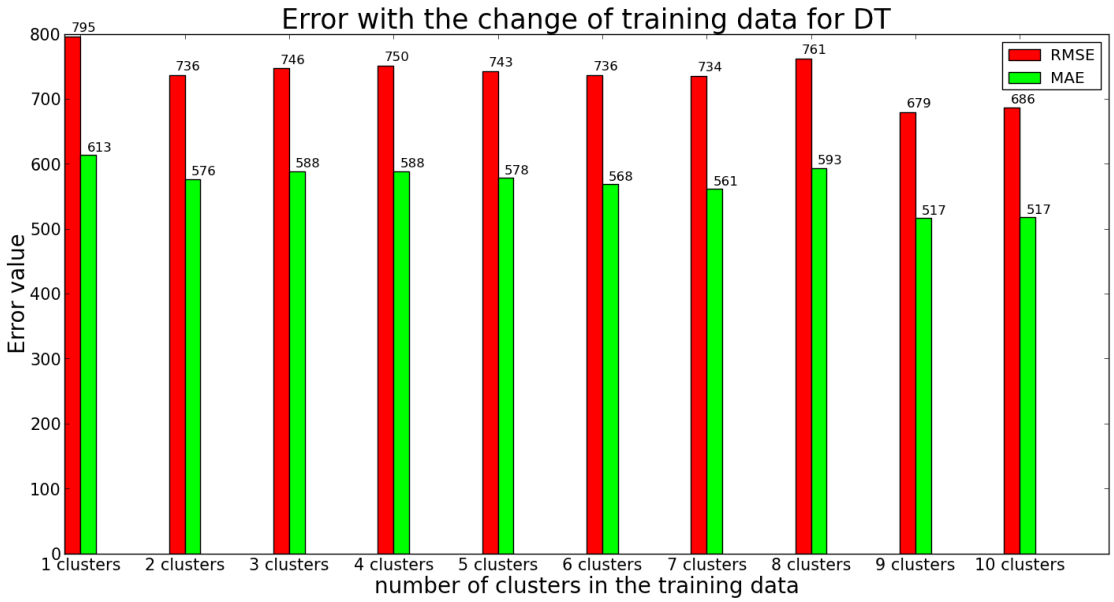

Figure 18: This figure shows the change of prediction error of DT over the process of appending one cluster of data from the training data pool to training data each time for building AHCMs.

Figure 19: This figure shows the change of prediction error of RaF over the process of appending one cluster of data from the training data pool to training data each time for building AHCMs.

Figure 20: This figure shows the change of prediction error of GBR over the process of appending one cluster of data from the training data pool to training data each time for building AHCMs.

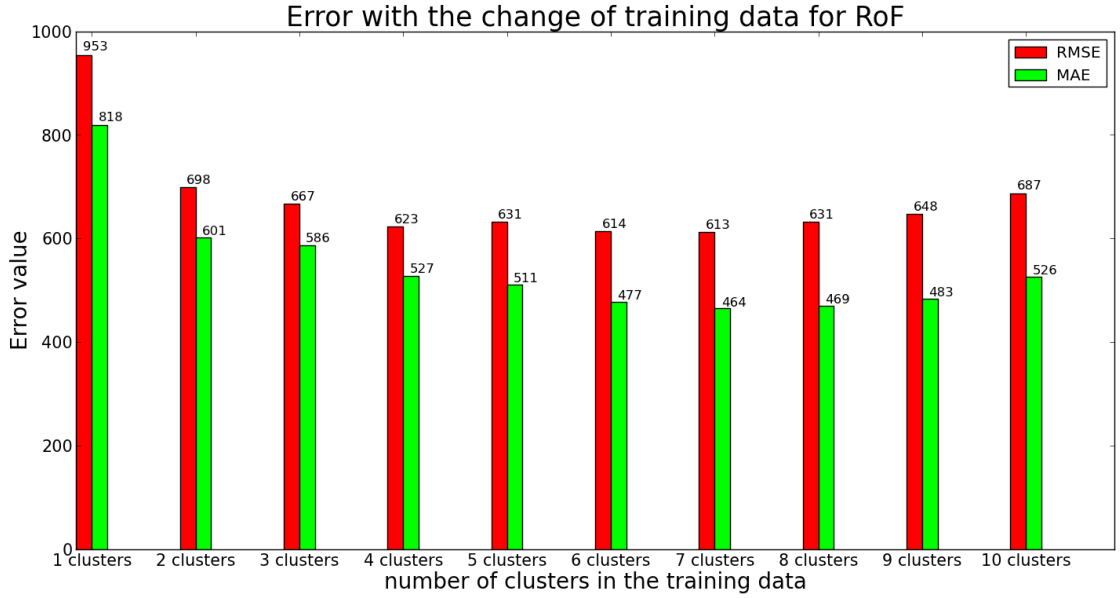

Figure 21: This figure shows the change of prediction error of RoF over the process of appending one cluster of data from the training data pool to training data each time for building AHCMs.

training data is 9. The trend of prediction error of RaF, GBR and RoF as shown in Figure 19 to Figure 21 is almost the same as them of HCMs individually. The best performance of them is shown in Table 10.

machine learning models	number of clusters in training data	RMSE	MAE
OLS	10	565	430
DT	9	679	517
RaF •	9 •	515 •	392 •
GBR	10	577	495
RoF	7	613	464

Table 10: This table shows the best prediction result and the best parameters for selecting the number of clusters in the pool of training data for five machine learning algorithms individually.

machine learning models	HCMs	AHCMs
OLS	0.8%	0%
DT	0%	1%
RaF	0.4%	7%
GBR	4%	0%
RoF	16%	11%

Table 11: This table shows improvement of the best HCMs and AHCMs compared to the best global model when using five different machine learning models

For AHCMs with DT and RaF, the performance of them has improved compared to the result of HCMs with DT and RaF. RMSE of the best AHCM with RaF is 6% less than the best HCM with RaF. In addition, it is also the best model obtained compared to other local models and global models although the improvement is only 7%. Moreover, the number of clusters involved in the best AHCM is 9 as shown in the table. However, the best AHCMs with OLS, GBR and RoF is worse than the best HCMs with them respectively. The worst one is AHCM with RoF as RMSE of it is 6% larger than the best HCM with RoF. Table 11 shows the improvement of AHCMs and HCMs compared to global models when different models are chosen.

It is obvious that the AHCM and HCM can improve prediction accuracy indeed but the highest improvement is only 16% with only changing number of clusters in the training data. Therefore, our proposed algorithms for selecting training data to build local models: HCMs, MHCMs and AHCMs can indeed improve performance of predictive models but the contribution to improving performance is not that significant since the highest improvement this algorithms can achieve is 16%.

Figure 22 reveals the performance of the best HCM, AHCM and global model on each cluster on testing data. Since all data points in Africa are testing data and Africa data consists of 8 clusters: cluster 1 to 7 and cluster 9. Figure 22 shows RMSE of those three models on each cluster in Africa. Both the HCM and AHCM improve performance of the global model on cluster 2 and 5. The highest improvement is the AHCM on cluster 5. The improvement is around 12%. In addition, the global model has extraordinary performance on cluster 4 and it is 41% better than the best HCM and 20% better than the AHCM.

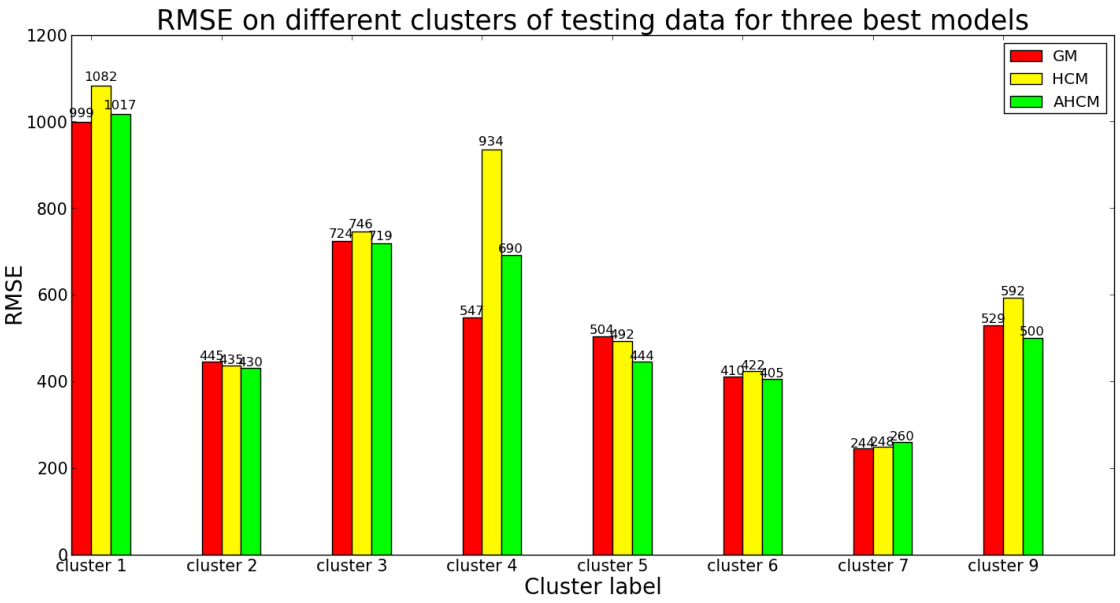

Figure 22: This figure shows performance of the best global model, HCM and AHCM on each clusters in testing data.

5.2 Results of models after tuning parameters

Results of models before tuning parameters as illustrated in section 5.1 prove that selecting data points based on similarity as training data can improve performance indeed. However, the performance of all models in section 5.1 can be improved by tuning parameters of machine learning models by using validation data. Thus, in this section, results of global models and local models after tuning parameters are described and the optimal parameters result in the best performance of models are illuminated as well. Thus, explanation of prediction results of models after tuning parameters starts from global models as following.

5.2.1 Global models

Table 12 shows RMSE and MAE on three test folds. As mentioned in previous section, number of data points on three test folds are the same. Three optimal global models are built by using GBR with tuned parameters for three test folds respectively. Among five machine learning algorithms: OLS, DT, RaF, GBR and RoF, GBR are selected for each test fold. In addition, the optimal global model on test fold 2 has the best performance. Finally, RMSE and MAE of global models on the whole Africa data which consists of test fold 1, 2 and 3 are 488 and 384. The performance of global models after tuning parameters improved 12% compared to the best global model before tuning parameters.

	RMSE	MAE	Optimal machine learning algorithms	parameter settings
Test fold 1	588	481	GBR	estimators: 4, learning rate: 0.3058, depth: 1
Test fold 2 •	381 •	295 •	GBR •	estimators: 6, learning rate: 0.481, depth: 1 •
Test fold 3	472	377	GBR	estimators: 6, learning rate: 0.481, depth: 1

Table 12: This table shows the best machine learning algorithm for global models after tuning parameters and prediction error of three test folds

5.2.2 Baseline models

From this section to section 5.2.6, results of five proposed local models after tuning parameters of machine learning algorithms are illustrated. This section illuminates the prediction results of baseline models. Table 13 shows RMSE and MAE of baseline models after tuning parameters on three test folds. RMSE and MAE of baseline models on the whole testing data means RMSE and MAE are calculated based on the prediction made on three test folds.

	RMSE	MAE
Test fold 1	494	382
Test fold 2	606	448
Test fold 3	786	626
All testing data $\star$	652 $\star$	495 $\star$

Table 13: This table shows prediction results of baseline models after tuning parameters on three test folds and RMSE and MAE of predictions over the whole testing data

Table 13 reveals that performance of the baseline model on test fold 1 is the best compared to the rest two folds. Test fold 1 is a group of data points on the west of the Africa continent and data points of test fold 2 are located in the center area of Africa continent; test fold 3 is the east part of Africa including Madagascar. Thus, the trend of performance of baseline models from west to east is that data points tend to more and more difficult to predict. RMSE on test fold 3 is 59% larger than that of test fold 1. RMSE over the whole Africa is 652 and it is 11.5% less than RMSE of the best baseline model before tuning parameters. Thus, tuning parameters can improve performance of baseline models indeed. But, it is still 18% larger than RMSE of the best global models. Thus contribution of tuning parameters for improving baseline models performance is limited if training data selected are not enough.

Table 14 gives optimal algorithms and its best parameters settings after tuning parameters by using validation data for each layers on three test folds. GBR are selected as an optimal algorithms in most situations of baseline models. Figure 23 shows the same results as Figure 10 but two lines in the figure represent RMSE and MAE of baseline models after tuning parameters and training data and testing data

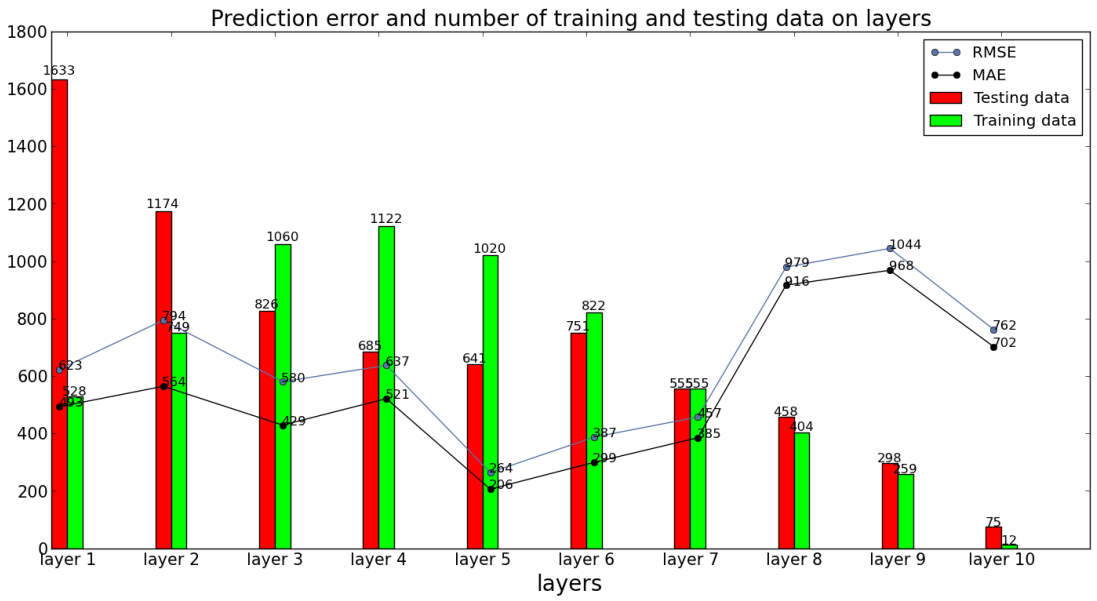

Figure 23: This figure shows RMSE and MAE on each layer for baseline models after tuning parameters and the change of number of training data and testing data over different layers

are the same as baseline models before tuning parameters on each layer. Compared RMSE of layer 1 and layer 10 of baseline models before tuning parameters with RMSE of the same two layers of baseline models after tuning parameters, performance for both two layers are around 1% worse in Figure 23. However, performance of baseline models on rest of layers are improved after tuning parameters as shown on Figure 23. Tuning parameters of baseline models contribute 55% improvement of performance on layer 5. Assuming that equator is the criteria for seperating north and south of Africa, tuning parameters of baseline models contribute more on the south of Africa as equator is located in layer 3.

5.2.3 Modified baseline models

Prediction results of modified baseline models after tuning parameters are illuminated in this section. Like in section 5.2.2, Table 15 shows RMSE and MAE of modified baseline models with best tuned parameters and machine learning models on three test folds. In addition, the row marked with star represent RMSE and MAE of modified baseline models on the whole testing data. Moreover, performance

	Test fold 1	Test fold 2	Test fold 3
layer 1	GBR(16, 0.471, 3)	GBR(24, 0.356, 2)	GBR(7, 0.296, 2)
layer 2	GBR(11, 0.351, 1)	GBR(11, 0.351, 1)	GBR(13, 0.316, 2)
layer 3	GBR(25, 0.331, 3)	GBR(21, 0.501, 2)	RaF(17)
layer 4	GBR(2, 0.011, 1)	GBR(1, 0.0001, 3)	GBR(1, 0.301, 1)
layer 5	DT(2)	DT(2)	GBR(25, 0.361, 2)
layer 6	RaF(12)	RaF(9)	GBR(14, 0.481, 2)
layer 7	GBR(25, 0.386, 2)	GBR(15, 0.271, 1)	GBR(18, 0.446, 2)
layer 8	GBR(1, 0.0001, 1)	GBR(1, 0.0001, 1)	GBR(1, 0.0001, 1)
layer 9	GBR(1, 0.0001, 3)	GBR(1, 0.0001, 3)	GBR(1, 0.0001, 3)
layer 10	RaF(4)	RaF(4)	RaF(5)

Table 14: This table shows the best machine learning algorithm and its optimal parameter settings for baseline models on different layers for three test folds

	RMSE	MAE
Test fold 1	425	347
Test fold 2	360	274
Test fold 3	718	571
All testing data $\star$	533 $\star$	402 $\star$

Table 15: This table shows prediction results of modified baseline models on three test folds and RMSE and MAE of predictions over the whole testing data

of modified baseline models after tuning parameters on test fold 2 is the best with RMSE 360, compared to test fold 1 and 2. In addition, performance of modified baseline models on three test folds are all improved, compared with prediction result of best tuned baseline models on three test folds. More importantly, RMSE on test fold 2 is 41% less than the result of best tuned baseline models on test fold 2. Thus best tuned modified baseline models contribute mainly on improving performance on test fold 2. In this case, data points in test fold 1 and test fold 3 are more difficult to predict compared to data points in test fold 2. As for overall performance of fine tuned modified baseline models on the whole testing data, it is also improved 18% compared to the same type of result of best tuned baseline models. Although overall performance of best tuned modified baseline models is better than best tuned baseline models, its performance on data points in layer 3 and layer 5 are 8% and 17.5% worse than performance of best tuned baseline models on the same layers as

shown in Figure 24. Therefore, if performance of a predictive model is better than another model on the whole data points, the later one can still have possibility that it can have relatively better results on a small part of data points.

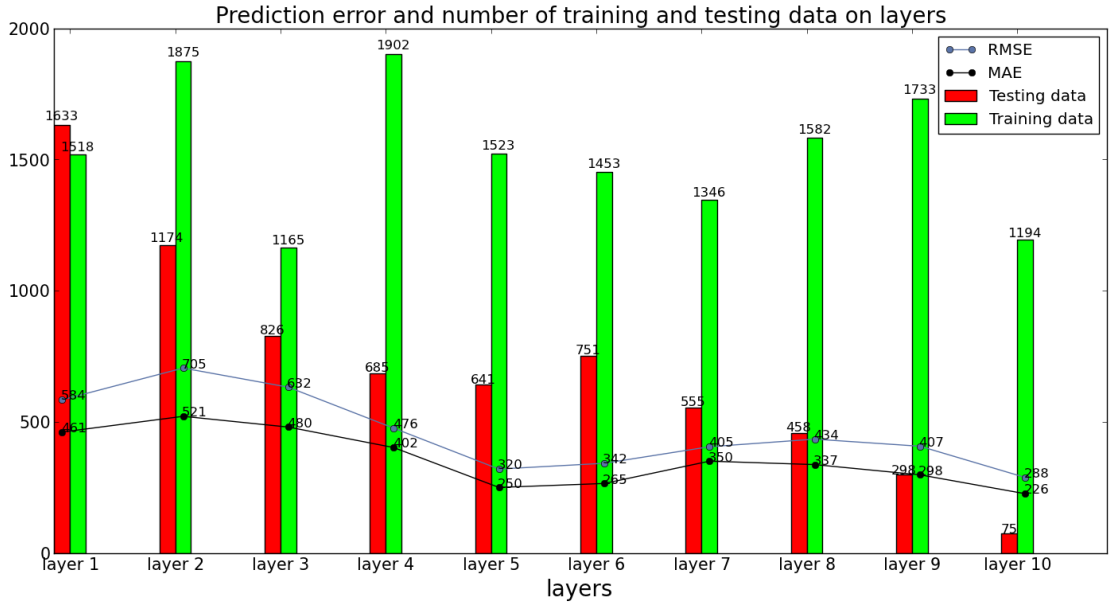

Figure 24: This figure shows RMSE and MAE on each layer for the best tuned modified baseline models and the change of number of training data and testing data over different layers

Furthermore, Figure 24 also reveals the same trend that data points in the south is harder to predict compare the north of equator. For best tuned modified baseline models, prediction result on the $10th$ layer has the best RMSE which is 288 and it is the south most layer. Moreover, RMSE of the best tuned modified baseline models on the whole testing data is 4% less than modified baseline model before tuning parameters. Moreover, performance of the best tuned modified baseline models on all testing data has improved 28% of the best baseline model before tuning parameters. Thus, 4% of improvement is contributed by parameters tuning and 24% of improvement is contributed by appending more training data for building modified baseline models, compared performance of the best tuned modified baseline models with baseline models without tuning parameters. Thus in the case of comparing baseline models and modified baseline models, the contribution of improving models performance by appending more closely matched data points of testing data is 20% larger than the way of tuning parameters. In addition, Table 16 shows parameters

settings and machine learning algorithms selected for all layers on three test folds separately. The table shows that GBR is most frequently chosen as the best machine learning models. Finally, the overall RMSE on all testing data is still worse than the best tuned global models with RMSE 488.

	Test fold 1	Test fold 2	Test fold 3
layer 1	GBR(25, 0.381, 4)	RaF(6)	GBR(12, 0.191, 2)
layer 2	RaF(1)	GBR(14, 0.366, 2)	GBR(3, 0.501, 2)
layer 3	GBR(20, 0.376, 3)	GBR(21, 0.396, 3)	GBR(4, 0.246, 4)
layer 4	GBR(21, 0.456, 2)	GBR(21, 0.466, 2)	GBR(23, 0.326, 2)
layer 5	GBR(11, 0.156, 2)	RaF(5)	GBR(10, 0.451, 3)
layer 6	GBR(18, 0.311, 3)	GBR(16, 0.391, 4)	GBR(25, 0.406, 3)
layer 7	GBR(6, 0.451, 2)	GBR(9, 0.396, 1)	GBR(10, 0.491, 3)
layer 8	GBR(5, 0.471, 4)	GBR(21, 0.451, 4)	DT(2)
layer 9	GBR(1, 0.236, 3)	GBR(4, 0.371, 4)	GBR(4, 0.371, 4)
layer 10	OLS	OLS	GBR(24, 0.311, 2)

Table 16: This table shows the best machine learning algorithm and its optimal parameter settings for modified baseline models on 10 different layers for three test folds

5.2.4 Hierarchical clustering based models

In this section, prediction results of fine tuned hierarchical clustering based models(HCM) are analysed. In the Figure 25, it shows the distribution of 8 clusters on the Africa continent. As mentioned in previous sections, Africa data points consist of 8 clusters: cluster 1 to 7 and cluster 9. As shown in the figure, the majority of data points in cluster 1 are located in warm semi−arid climate zone. Data points in cluster 2 are located in climate zone: equatorial climate zone and monsoon climate zone. Almost all data points in cluster 3 and 4 are located in Madagascar. Cluster 5 can represent tropical savanna climate zone and subtropical climate zone. Cluster 6 and 7 can stand for desert climate zone. Finally, data points of cluster 9 are located in part of cold semi−arid climate zone and warm semi−arid climate zone.

Table 17 shows the prediction result of HCM on three test folds separately. Performance of best tuned HCM on test fold 2 which is the center part of Africa is

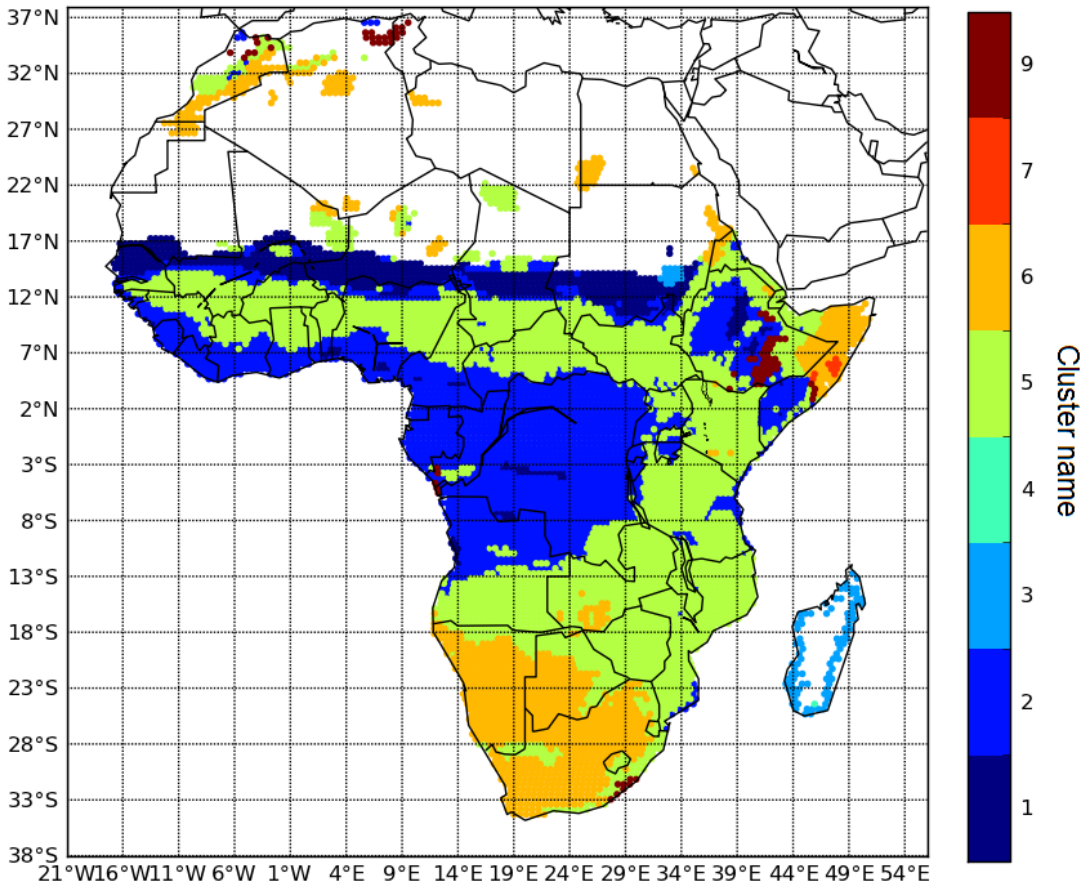

Figure 25: This figure shows the location of 8 clusters on the Africa continent.

the best with RMSE 378. In addition, RMSE of best tuned HCM on test fold 1 is 14% smaller than RMSE of the HCM on test fold 3. Performance of the best tuned HCM on test fold 3 which is the eastern part of Africa is the worst compared to test fold 1 and 2. Meanwhile, performance of the best tuned HCMs on the whole testing data is improved 12% of the best tuned global model. Compared performance of it with global models without tuning parameters with RMSE 552, the performance is improved 22.3%. Thus tuning parameters of machine learning parameters and selecting data points as training data that can match testing data closely together can improved general performance of models significantly on the whole testing data. Moreover, it is also improved 19.5% of the best tuned modified baseline models. Thus, in the perspective of general performance, the best tuned HCMs is the best model compared all models mentioned in previous sections. Table 18 shows pa-

rameters settings of the best machine learning models for each cluster on each test fold and the number of clusters from the training data pool are selected. Table 19 shows vectors of rank of clusters in training data pool for 8 clusters in Africa. For example, the first row in table 19 means for cluster 1, the rank of clusters based on similarity is $\{1, 2, 3, 4, 7, 8, 9, 10, 5, 6\}$. As shown in table 18, in the first row, for cluster 1 in Africa, in the process of building the best tuned HCM, GBR with number of estimators: 1, learning rate: 0.0001 and number of depth 1 and 10 clusters in the training data pool which are data points in Eurasia, North America and South America are selected as training data. Therefore, the best tuned HCM for cluster 1 is a global model since all data points in training data pool are selected for building a model. For instance, if selected number of clusters in training data is 5 for cluster 5 as shown in the 5th row of table 19, then selected training data consists of cluster $\{5, 6, 9, 10, 7\}$ of training data pool, which is first 5 clusters in the vector shown the rank of clusters of cluster 5 in Africa.

	RMSE	MAE
Test fold 1	417	337
Test fold 2	378	289
Test fold 3	486	378
All testing data $\star$	429 $\star$	335 $\star$

Table 17: This table shows prediction results of fine tuned HCM on three test folds and RMSE and MAE of predictions over the whole testing data

In order to understand performance of best tuned HCMs on different clusters, Figure 26 shows RMSE and MAE on 8 clusters separately. Height of red bars and green bars represent RMSE and MAE. Although it is not rigorous to conclude that performance of best tuned HCMs on a cluster is better than its performance on another clustered based on their RMSE as the number of data points is different in different clusters, RMSE can at least provide a trend to indicate which cluster of testing data is easy to predict. Thus, cluster 2 is a part of testing data that are easiest to predict with minimum RMSE 348 and it is place where climate is relatively humid and there are forests. In addition, cluster 9 is the most difficult to predict with RMSE 678. Data points of cluster 9 are located in the South east corner of Ethiopia, top corner of boundaries of Algeria and Tunisia and the top north corner of Morocco. Data points on Madagascar are the second most difficult to predict since its RMSE is 638.

	Test fold 1		Test fold 2		Test fold 3	
Cluster	Optimal models	♯clusters	Optimal models	♯clusters	Optimal models	♯clusters
1	GBR$(1, 0.0001, 1)$	10	GBR$(1, 0.0001, 1)$	10	GBR$(1, 0.0001, 1)$	10
2	GBR$(7, 0.376, 2)$	3	GBR$(14, 0.501, 1)$	2	GBR$(17, 0.381, 1)$	2
3	–	–	–	–	GBR$(22, 0.376, 3)$	5
4	–	–	–	–	RaF(1)	4
5	GBR$(25, 0.446, 4)$	9	GBR$(9, 0.391, 4)$	5	GBR$(22, 0.341, 2)$	1
6	GBR$(15, 0.166, 1)$	9	GBR$(5, 0.471, 3)$	2	GBR$(25, 0.161, 4)$	5
7	–	–	–	–	OLS	4
9	GBR$(3, 0.131, 4)$	7	GBR$(7, 0.081, 2)$	7	DT(15)	7

Table 18: This table shows the best machine learning algorithm and its optimal parameter settings for tuned HCMs and number of clusters in training data pool are selected as training data for three test folds

	rank of clusters
cluster 1	$\{1, 2, 3, 4, 7, 8, 9, 10, 5, 6\}$
cluster 2	$\{2, 1, 3, 4, 7, 8, 9, 10, 5, 6\}$
cluster 3	$\{3, 4, 1, 2, 9, 10, 7, 8, 5, 6\}$
cluster 4	$\{4, 3, 1, 2, 9, 10, 7, 8, 5, 6\}$
cluster 5	$\{5, 6, 9, 10, 7, 8, 3, 4, 1, 2\}$
cluster 6	$\{6, 5, 9, 10, 7, 8, 3, 4, 1, 2\}$
cluster 7	$\{7, 8, 9, 10, 5, 6, 1, 2, 3, 4\}$
cluster 9	$\{9, 10, 7, 8, 5, 6, 1, 2, 3, 4\}$

Table 19: This table shows vectors of clusters as ranks of clusters in training data pool for 8 clusters in testing data.

For desert area in Africa, performance of the best tuned HCMs on cluster 6 is much worse than that on cluster 7. But the number of data points in cluster 7 is only 9. In addition, data points in cluster 6 and cluster 5 have almost the same difficulty in making predictions.

In order to compare performance of the best tuned HCMs on different places in Africa strictly, data points on Africa are partitioned into 15 different slides vertically from the West to the east and each slides has the same number of data points, which is 549. In addition, the whole testing data are also divided into 15 different slides with equal number of data from the south to the north of Africa. Moreover,

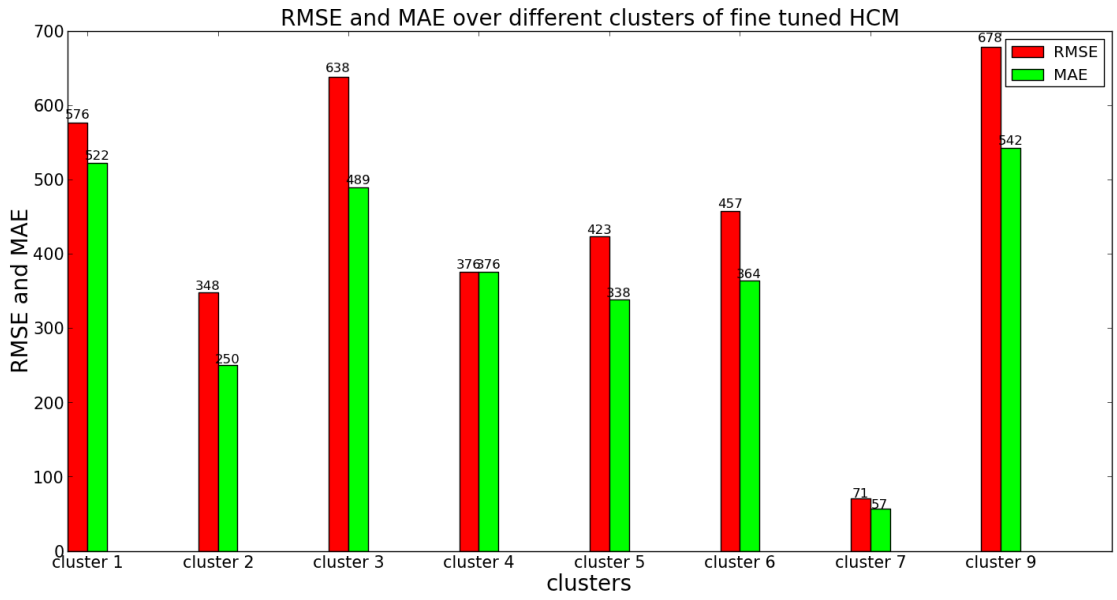

Figure 26: This figure shows RMSE and MAE on different clusters for the best tuned HCMs

RMSE and MAE are calculated in all layers based on the prediction results. Then, we mark data points in a layer with the same RMSE and project their locations on the map for the purpose of visualizing the change of RMSE over layers on the Africa continent. RMSE and MAE of the models on different slides from south to the north of Africa is shown in Figure 27 and Figure 28 represent RMSE and MAE on different slides from south to north projected to the map of Africa. Figure 29 and Figure 30 stand for the same result for different slides from the west to east of Africa. Moreover, color maps in Figure 28 and Figure 30 represent the value of RMSE and different color on the map shows different value of RMSE. For example, if color of data points on the map is red, their RMSE are around 320. Heights of red bars and green bars in Figure 27 and Figure 29 represent RMSE and MAE and layers means slides. For example, layer 1 in Figure 27 is the south most slide of Africa with number of data points 549 and layer 10 is the most north slide of Africa. Likewise, layer 1 in Figure 29 is the most West slide of Africa and layer 10 is the most east slide of Africa.

As shown in Figure 27 and Figure 28, among 15 layers, RMSE and MAE of the best tuned HCMs is the smallest on layer 6 and RMSE and MAE is the largest on layer 14; the smallest error is 52% less than the largest error. RMSE of tuned HCMs on

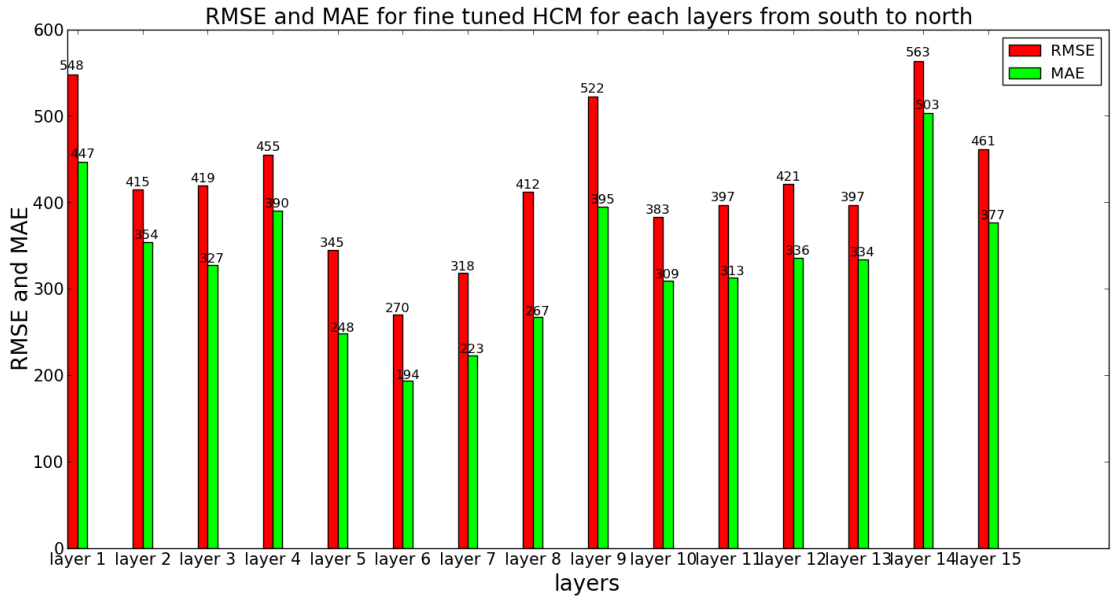

Figure 27: This figure shows RMSE and MAE of tuned HCMs on each layer from the south to the north of Africa

layer 1, layer 9 and layer 14 are larger than 500. As shown in Figure 28, layer 1 and layer 14 covers desert climate zone in the south corner of Africa. layer 9 covers Turkana Basin where fossil data points are located. More importantly, equator is located in layer 7. Therefore, in general, performance of tuned HCMs on Africa from South to North is hemispheric symmetry. Therefore, from layer 1 to layer 6, RMSE of tuned HCMs decreases gradually and from layer 7 to 15, RMSE increases slowly. Furthermore, the trend of prediction error of tuned HCMs from west to east of Africa is that from layer 1 to layer 2, RMSE is decreased 11%; From layer 2 to layer 4, RMSE increased from 405 to 444 and RMSE of layer 4 is one of maxima among RMSE of all layers. Then from layer 4 to layer 5, RMSE decreases sharply and it is decreased 22.5%. From layer 5 to layer 8, RMSE remains almost the same which is around 344 and performance on layer 8 is the best with RMSE 343. Then, RMSE increases from layer 8 to layer 11 and it increased 45%. Finally, RMSE decreases from layer 11 to layer 12 and it start increasing until layer 15 and RMSE on layer 15 is the worst which is 608. The smallest RMSE which is on layer 8 is only 57% of the largest RMSE. Thus, data points in the west part of Africa from layer 9 to layer 15 is difficult to predict than the east part of Africa.

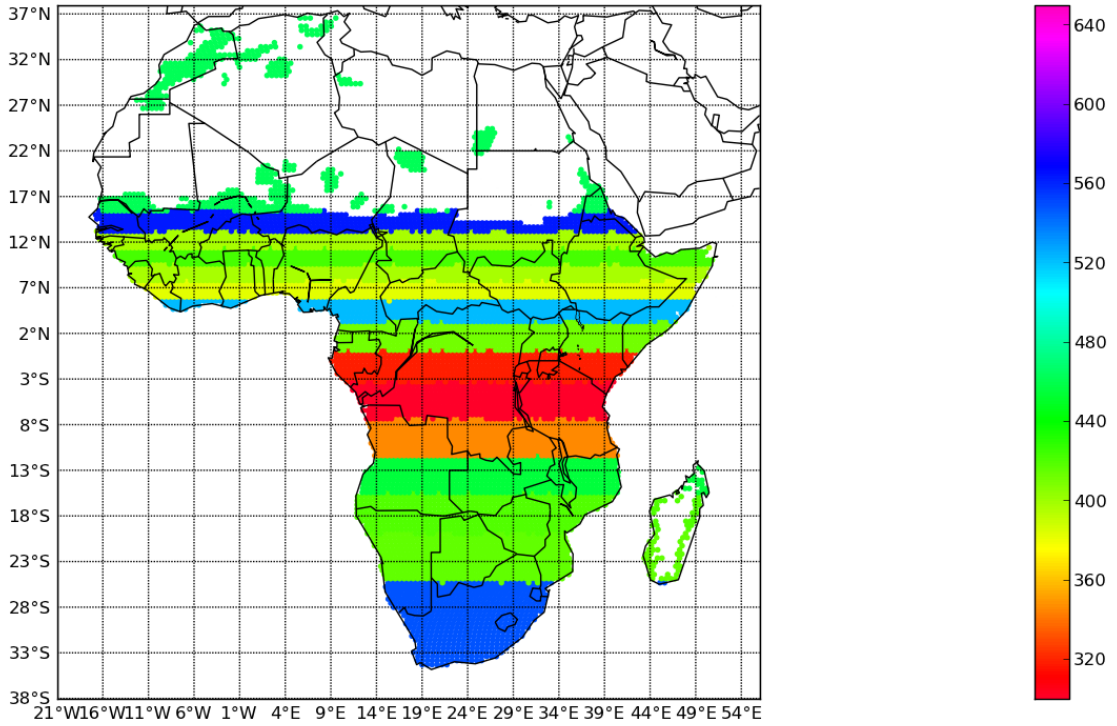

Figure 28: This figure shows RMSE of tuned HCMs on each layer from the south to the north and map location of layers to Africa continent on the world map. Data points in the same layer are marked with the same RMSE for the purpose of showing its location. This applies to similar figures showing RMSE on layers.

5.2.5 Modified hierarchical clustering based models

	RMSE	MAE
Test fold 1	550	368
Test fold 2	396	300
Test fold 3	526	399
All testing data ⋆	491 ⋆	355 ⋆

Table 20: This table shows prediction results of fine tuned MHCMs on three test folds and RMSE and MAE of predictions over the whole testing data

This section illuminates prediction results of fine tuned MHCMs. Table 20 illuminates RMSE and MAE of best tuned MHCMs on test folds and all testing data. Like the result in fine tuned HCMs, performance of fine tuned MHCMs on test fold 2 is the best compared to the other two test folds. In addition, RMSE of fine tuned

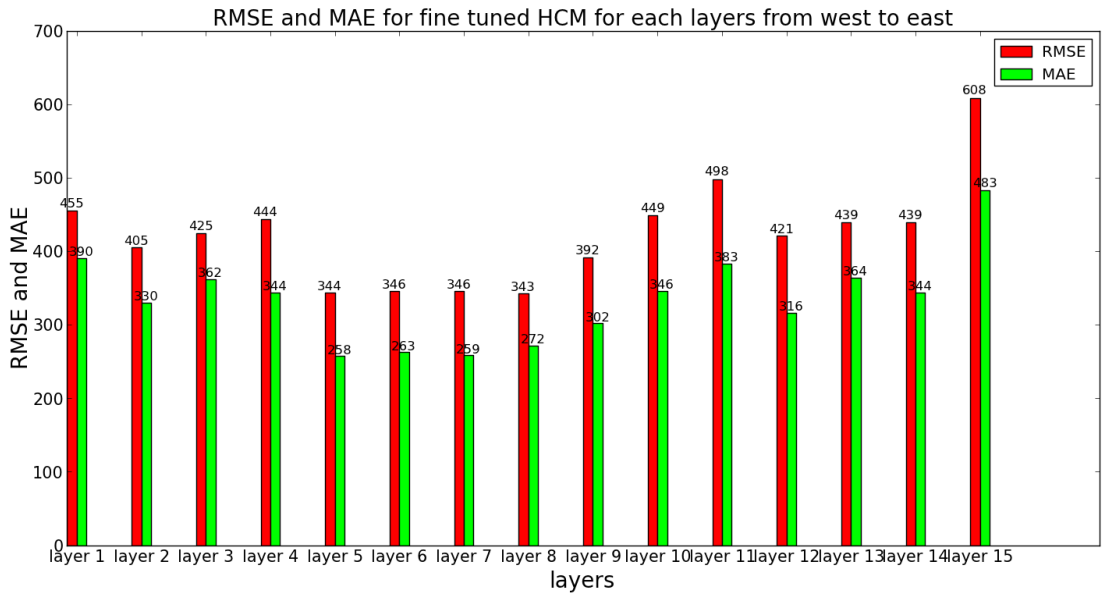

Figure 29: This figure shows RMSE and MAE of tuned HCMs on each layer from the west to the east of Africa

MHCMs on test fold 2 is 28% smaller than test fold 1. Moreover, RMSE of fine tuned MHCMs on three test folds are all worse than the result in table 17. Furthermore, performance of fine tuned MHCMs on test fold 1 is 32% worse than fine tuned HCMs. As for performance on the whole testing data, it is 14.5% worse than fine tuned HCMs and it is even 0.6% worse than fine tuned global models. Thus, according to RMSE and MAE on test folds or the whole Africa data, performance of fine tuned MHCMs is the worst compared to find tuned global models and HCMs. It is possible that fine tuned MHCMs is so flexible that it can overfitting since clusters with large number of data points are divided into several horizontal layers so that the number of data points in a layer is smaller than 500 and each layer in validation data are utilised to tune parameters of machine learning algorithms for building tuned MHCM.

Furthermore, Figure 31 shows performance of fine tuned MHCMs on 8 different clusters. Like in Figure 26, red bars and green bars represent RMSE and MAE value. Compared the result in Figure 31 with Figure 26, performance of fine tuned MHCMs in cluster 1, 2, 5 and 6 are all worse than it of fine tuned HCMs. Performances of the MHCMs of the rest clusters are the same as HCMs. Since only data points in cluster 1, 2, 5 and 6 are tested with fine tuned MHCMs. For RMSE of

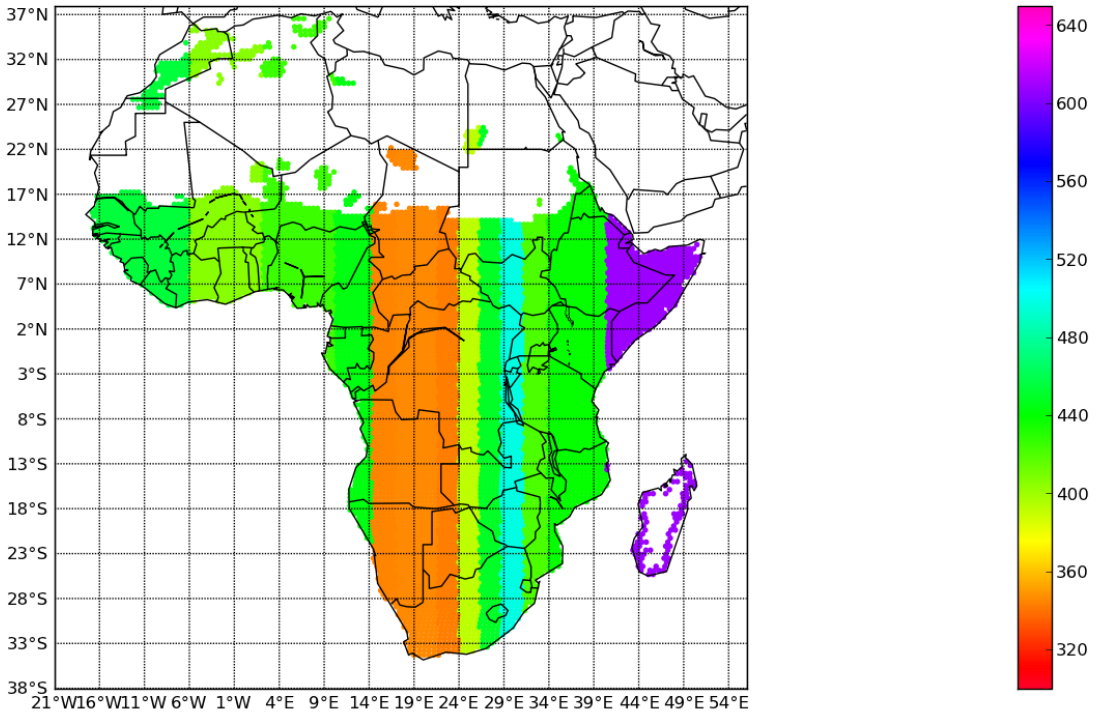

Figure 30: This figure shows RMSE of tuned HCMs on each layer from the west to the east and map location of layers to Africa continent on the world map

tuned MHCMs on cluster 1 is 34% worse than tuned HCMs on the same cluster. Thus, in the aspect of performance on different clusters respectively, MHCMs have no contribution in improving performance of HCMs on all clusters and it even has worse prediction results. Meanwhile, the trend of performance on different clusters is the same as HCMs like illustrating in section 5.2.4.

Like in Figure 27 to Figure 30, Figure 32 to Figure 35 shows the change of RMSE and MSE of tuned MHCMs over 15 slides with same number of data points from the south to the north of Africa and the West to the East of Africa. Compared the prediction result on 15 different horizontal layers of MHCMs, as shown in Figure 32 and Figure 33, with HCMs on the same layers, performance on layer 1, 5, 10 and 11 are all improved. In addition, the largest improvement appears in layer 11 and RMSE of tuned MHCMs on layer 11 is 18% smaller than tuned HCMs. As for vertical layers from west to east of Africa as shown in Figure 34 and Figure 35, RMSE of MHCMs on layer 8 and layer 14 are all 2% less than HCMs. Thus, although performance of MHCMs on the whole testing data or different clusters are

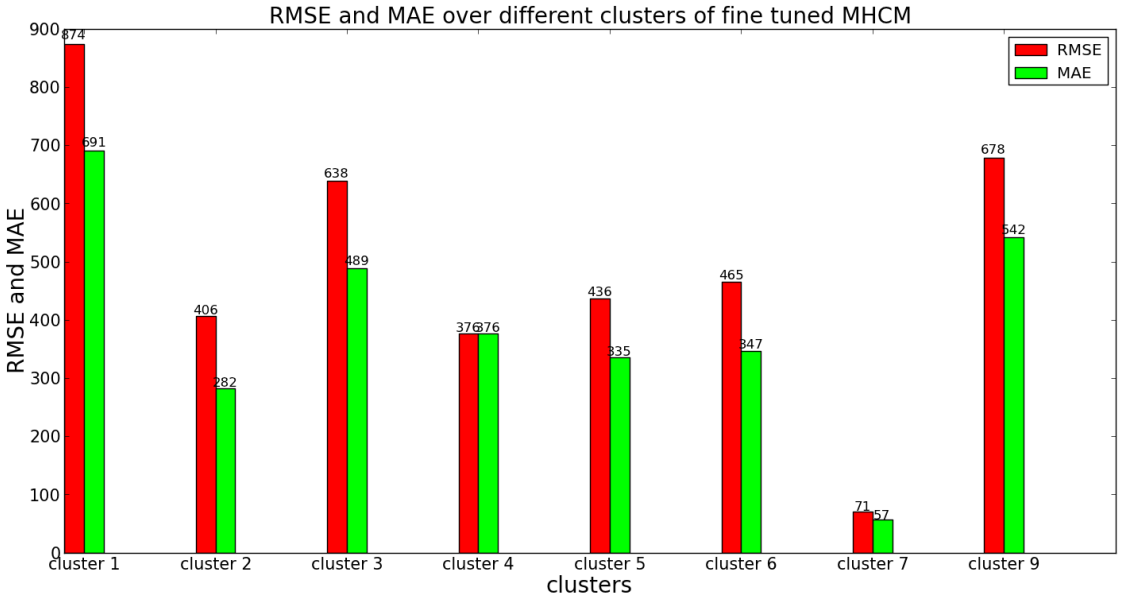

Figure 31: This figure shows performance of tuned MHCMs on 8 different clusters respectively.

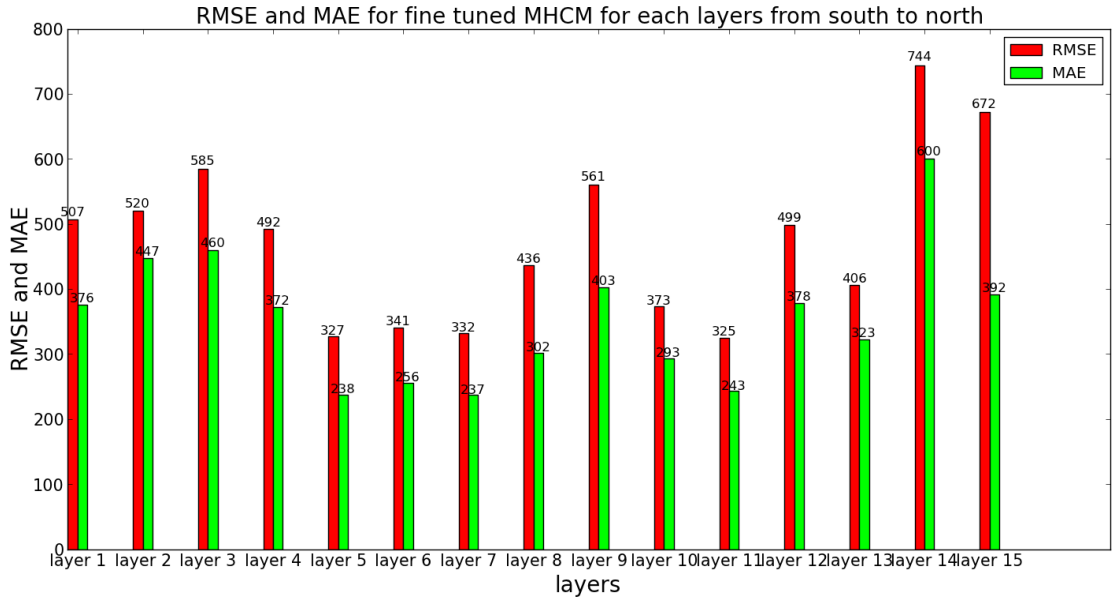

Figure 32: This figure shows performance of tuned MHCMs on 15 horizontal layers with equal number of data points from south to north of Africa.

worse compared to HCMs, MHCMs on some small parts of data points can have contribution in improving performance of HCMs, like horizontal layer 11. RMSE

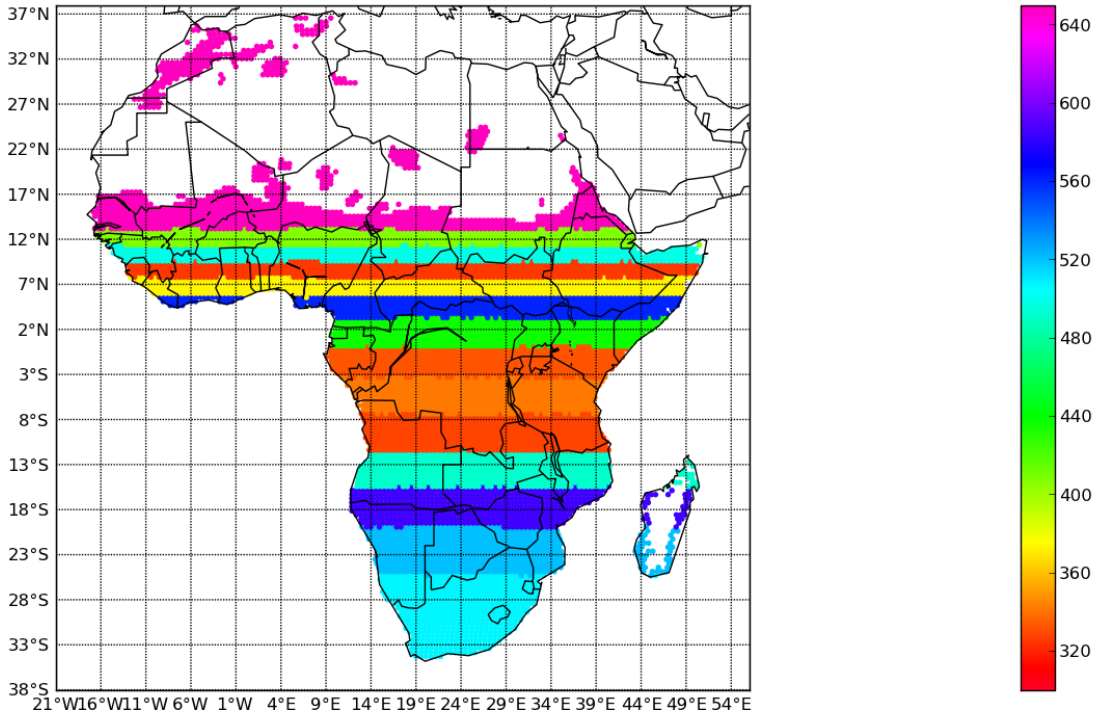

Figure 33: This figure shows RMSE of tuned MHCMs on each layer from the south to the north and map location of layers to Africa continent on the world map

of tuned MHCMs on them are much smaller than tuned HCMs. Furthermore, prediction results of tuned MHCMs as shown in Figure 33 reveals that data points that are located in the area of equator are easiest to predict and the performance of MHCMs on the north of equator and on the south of equator is symmetric. Moreover, as shown in Figure 35, performance of tuned MHCMs on the right side of vertical layer 8 and left side of layer vertical 8 are also symmetric and predictions of MHCMs on layer 8 has the best result and it is even 2% better than tuned HCMs.

5.2.6 Advanced hierarchical clustering based models

In this section, prediction results of fine tuned advanced hierarchical clustering based models(AHCMs) are illustrated. Moreover, comparison of fine tuned GMs, HCMs, MHCMs and AHCMs are also illuminated in this section and according to their performance on different clusters, a scheme of selecting models on different parts of Africa are also described.

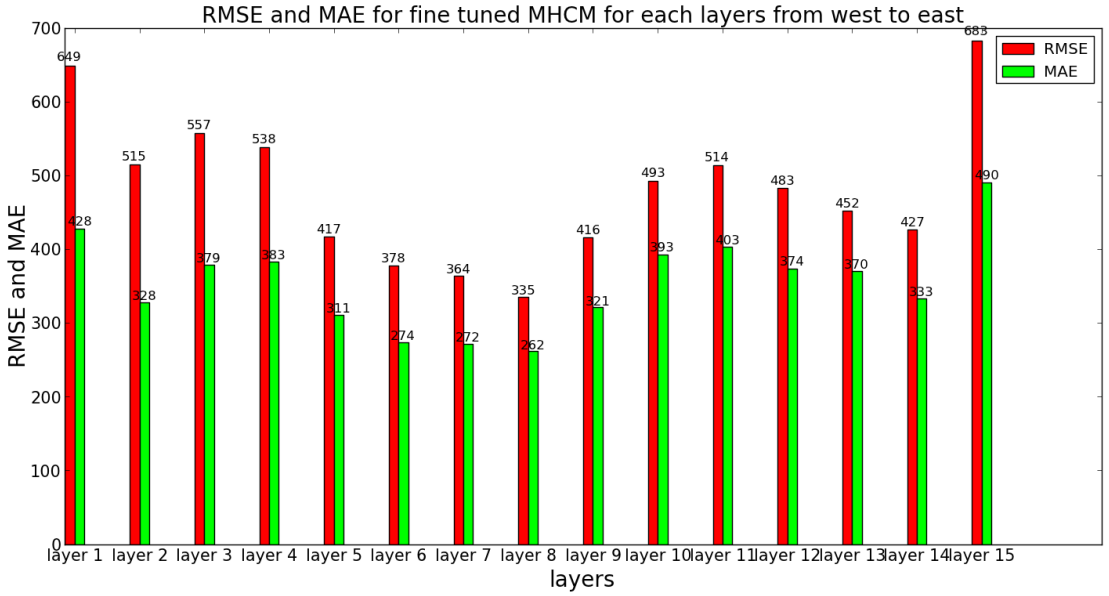

Figure 34: This figure shows performance of tuned MHCMs on 15 horizontal layers with equal number of data points from west to east Africa.

	RMSE	MAE
Test fold 1	388	300
Test fold 2	322	247
Test fold 3	425	327
All testing data $\star$	380 $\star$	291 $\star$

Table 21: This table shows prediction results of fine tuned AHCMs on three test folds and RMSE and MAE of predictions over the whole testing data

Firstly, Table 21 shows RMSE and MAE of fine tuned AHCMs on three test folds and the whole testing data. Moreover, RMSE of fine tuned AHCM on test fold 1 is the smallest among prediction results on three test folds and RMSE of the model on test fold 1 is 24.2% smaller than test fold 3. Thus from data points in test fold 1 to 3, data points in test fold 2 which are in the center part of Africa are the easiest to predict for tuned AHCMs and data points in test fold 1 which are located in the west part of Africa are the second easiest to make predictions for tuned AHCMs. Finally, performance of AHCMs on test fold 3 which are in the east of Africa is the worst. Furthermore, compared RMSE of tuned HCMs on three test folds with tuned

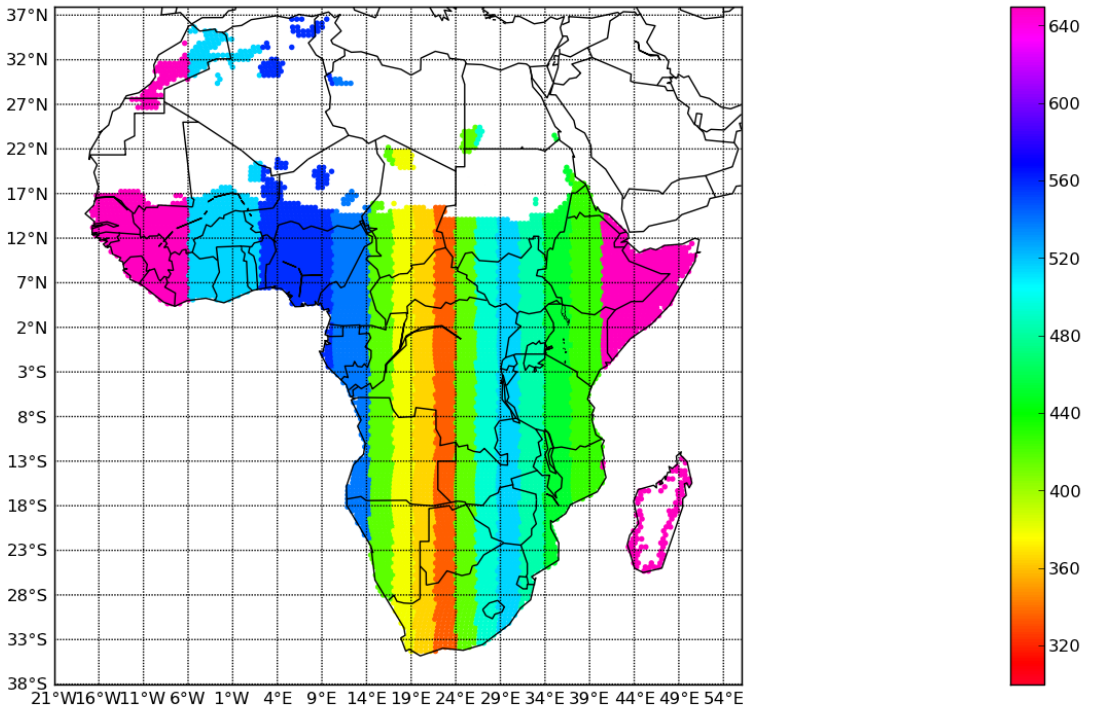

Figure 35: This figure shows RMSE of tuned MHCMs on each layer from the west to the east and map location of layers to Africa continent on the world map

AHCMs, performance of tuned AHCMs on each test fold is much more better. For test fold 1, RMSE of tuned AHCM is 7% better than tuned HCMs; For test fold 2, RMSE of AHCMs is 15% improved compared to HCMs on test fold 2; RMSE of AHCMs on test fold 3 is 13% smaller than HCMs. In addition, performance of tuned HCMs on three test folds are better than performance of tuned GMs and MHCMs on three test folds individually. Thus, in the perspective of individual performance of models on three test folds, tuned AHCMs are the best model among global models and local models described in this thesis. Ultimately, performance of tuned AHCMs on the whole testing data is improved 11.4% of tuned HCMs on the whole testing data and it is improved 39.2% of performance of baseline models on the whole testing data. Therefore, in the aspect of performance on the whole testing data, tuned AHCMs are the best among all global models and proposed local models.

In order to understand performance of tuned AHCMs on different clusters in Africa, RMSE and MAE of tuned AHCMs are calculated for each cluster. Moreover, RMSE and MAE of tuned global models(GMs), HCMs, MHCMs and AHCMs on each clus-

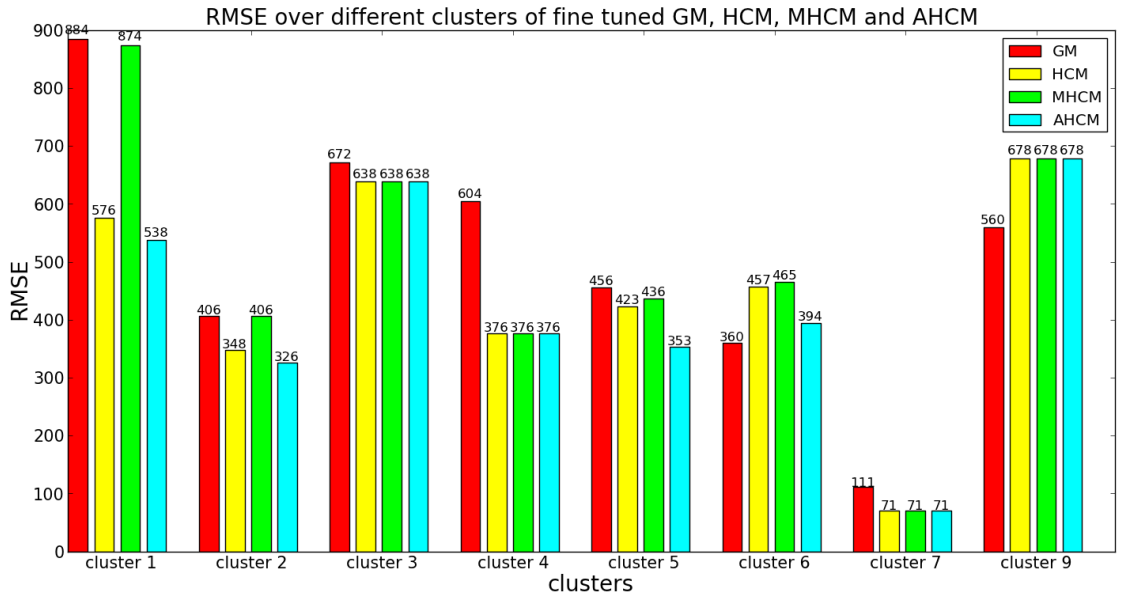

Figure 36: This figure shows RMSE of tuned GMs, HCMs, MHCMs and AHCMs on 8 clusters in Africa

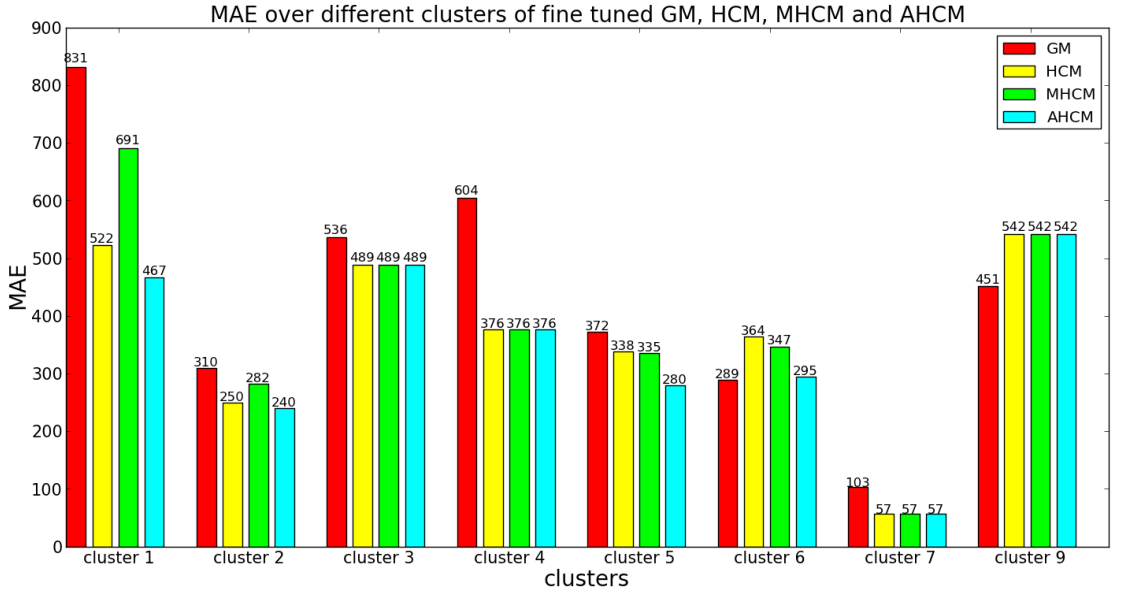

Figure 37: This figure shows MAE of tuned GMs, HCMs, MHCMs and AHCMs on 8 clusters in Africa

ter are shown in Figure 36 and Figure 37. In Figure 36 and Figure 37, red, yellow, green and light blue bars represent prediction error of those four models. Heights of bars in Figure 36 represent RMSE and heights of bars in Figure 37 stand for MAE.

In addition, tuned AHCMs have the best performance among three fine tuned clustering based models. For cluster 3, 4, 7 and 9, tuned HCMs, MHCMs and AHCMs have the same prediction result as numbers of data points in those clusters are all smaller than 104 so that those clusters are not suitable to partitioned into different layers like in MHCMs or clustering them into sub-clusters as in the process of building AHCMs. Therefore, predictions of MHCMs and AHCMs of those clusters are kept same predictions result of tuned HCMs. Moreover, As for general performance on cluster 1, 2, 5, 6, tuned AHCMs have the best performance and tuned HCMs are the second best local models; Tuned MHCMs are the worst clustering based local models. Moreover, tuned AHCMs contribute significantly to improve performance on data points in cluster 5 and RMSE of tuned AHCMs on cluster 5 is 16.5% smaller than tuned HCMs on cluster 5. Then, Compared performance of tuned global models with tuned AHCMs, performance of tuned global models in cluster 6 and 9 are better than performance of tuned AHCMs. RMSE of tuned global models on cluster 6 is 9% smaller than tuned AHCM and RMSE of tuned global models on cluster 9 is 17.4% smaller than tuned AHCMs on the same cluster. Furthermore, RMSE of tuned AHCMs on cluster 1, 4, 7 are all at least 36% smaller than tuned GM on those clusters. In addition, for cluster 2 and 5, RMSE of tuned AHCMs on those clusters are at least 20% smaller than tuned GM. Therefore, AHCMs improve performance of tuned GM significant on clusters 1, 2, 4, 5 and 7.

Figure 38 shows location of three sub-clusters of cluster 1 in Africa and Figure 39 shows performance of those four models on three sub-clusters of cluster 1. For subcluster 1 and 3 which have more plants than location of sub-cluster 2, performance of tuned AHCMs are still the best among those clusters. However, for sub-cluster 2, RMSE of tuned HCMs is 27% smaller than tuned AHCMs. Thus for sub-cluster 2, tuned HCM is the best model.

Figure 40 also shows distribution of 6 sub-clusters on African continent. Figure 41 also reveals performance of four models on 6 sub-clusters. For sub-cluster 1 in cluster 2, RMSE of tuned AHCMs is 37% smaller than tuned GMs. Thus, tuned AHCMs contributes much in improving performance of models on sub-cluster 1. Furthermore, prediction results of tuned HCMs are better than AHCMs on both sub-cluster 5 and 6. For sub-cluster 6, RMSE of tuned AHCMs is 20% smaller than tuned HCM. Subcluster 6 covers equatorial climate zone.

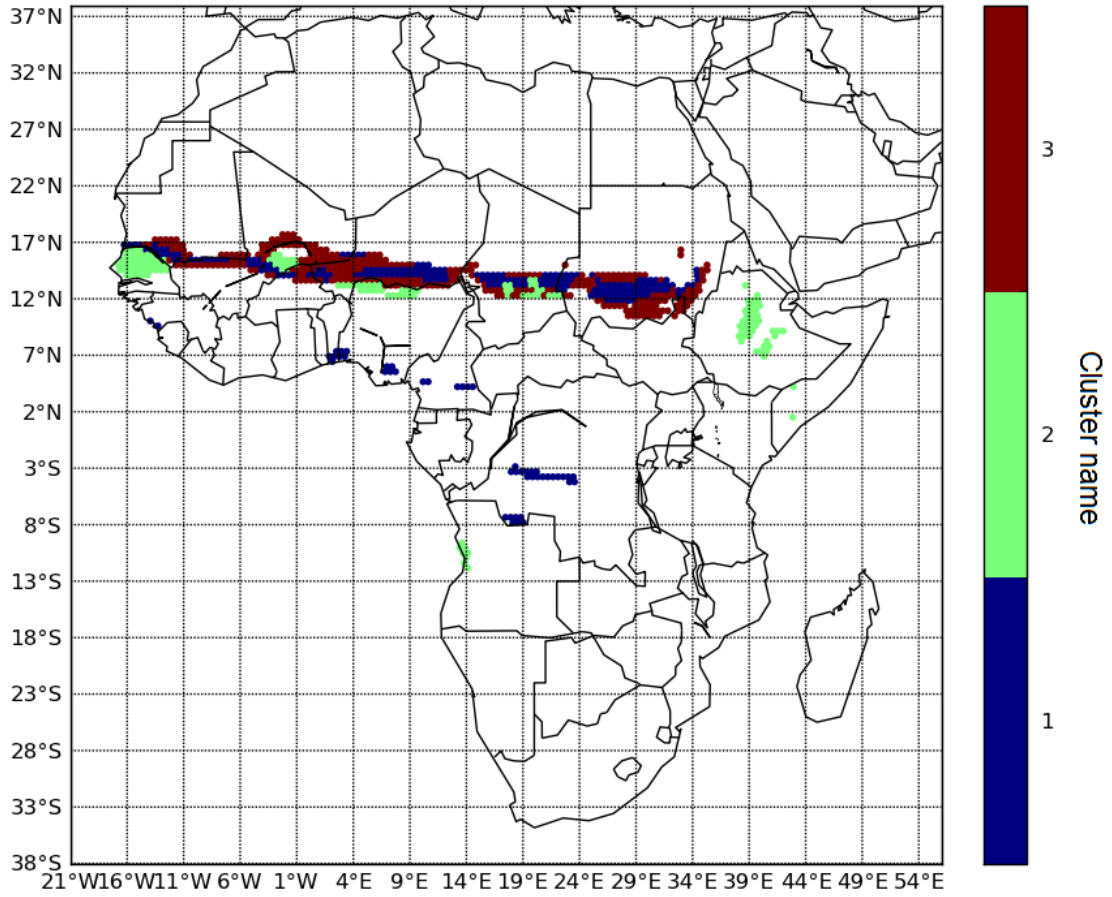

Figure 38: This figure shows cluster 1 with 3 sub-clusters on the map.

Figure 42 shows distribution of 10 sub-clusters of cluster 5 on Africa. As shown in Figure 43, tuned AHCMs contribute significantly in improving performance of global models on sub-cluster 1, 3 and 6. RMSE of tuned AHCMs on sub-cluster 1 is only 23% of of tuned global models on sub-cluster 1. Moreover, RMSE of tuned HCMs is 2.5% smaller than AHCM on sub-cluster 2. In addition, performance of tuned MHCM on sub-cluster 8 is 5% better than tuned AHCMs. Moreover, for sub-cluster 9 and 10, performance of tuned GM are all better than tuned AHCMs. Meanwhile, RMSE of tuned GMs on sub-cluster 10 is 11% better than tuned AHCMs. More importantly, sub-cluster 8, 9 and 10 covers Turkana Basin where fossil data are located. Therefore, For data points located in the area around Turkana lake, tuned MHCMs and GMs can have better prediction result.

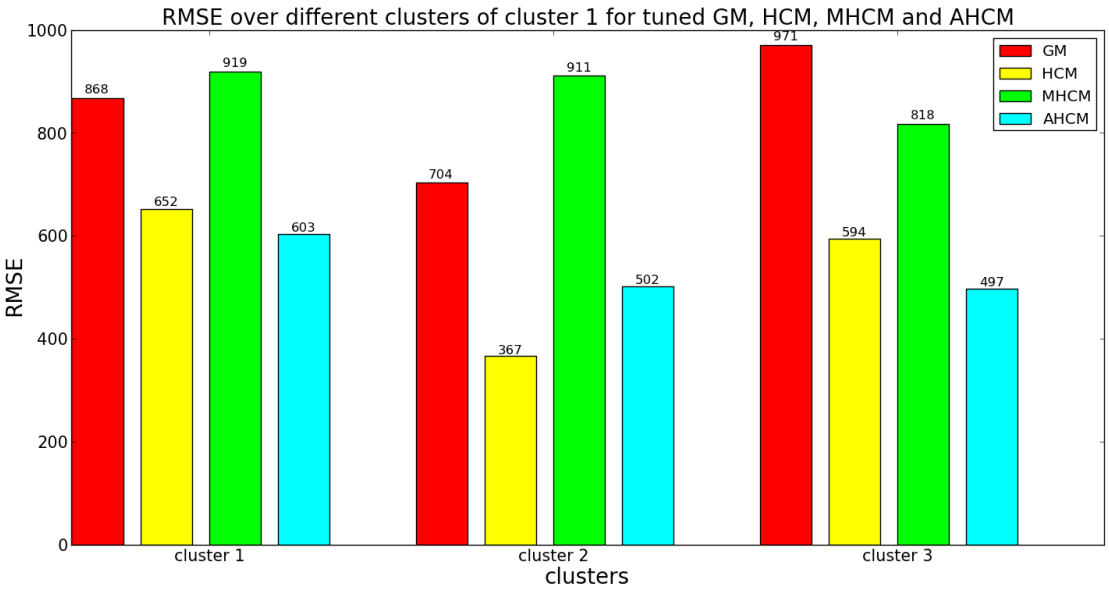

Figure 39: This figure shows RMSE of tuned GMs, HCMs, MHCMs and AHCMs for three sub-clusters of cluster 1 of Africa data

Figure 44 shows distribution of 4 sub-clusters of cluster 6 on Africa continent and as mentioned in previous paragraph in this section, tuned GMs has the best performance on cluster 6. However, as shown in Figure 45, performance of tuned HCMs on sub-cluster 3 is better than tuned GMs on sub-cluster 3. More importantly, tuned AHCMs can have better prediction result than tuned GMs on suclster 4 as well and RMSE of tuned AHCMs on this cluster is only 28% of RMSE of tuned GMs on the same cluster. Therefore, tuned HCMs and AHCMs can have better prediction result on the north part of cluster 6 and tuned GMs can have better result in mainly the south part of cluster 6.

Figure 46 shows RMSE of four models on 15 slides from the South to the North and Figure 47 shows RMSE of four models on 15 slides from the West to the East of Africa. The general trend is that data points in from horizontal layer 5 to layer 7 as shown in figure 46 are the easiest group of data points for making accurate predictions for all models. Likewise, for vertical layers from the West to the East of Africa, from data points in layer 5 to layer 8 are the group of data points that are very easy to predict. In addition, horizontal layer 9 and layer 14 in figure 46 are groups of data points that are the most difficult to predict for all models. More importantly, Turkana lake is also in layer 9. As shown in figure 46, tuned GMs can

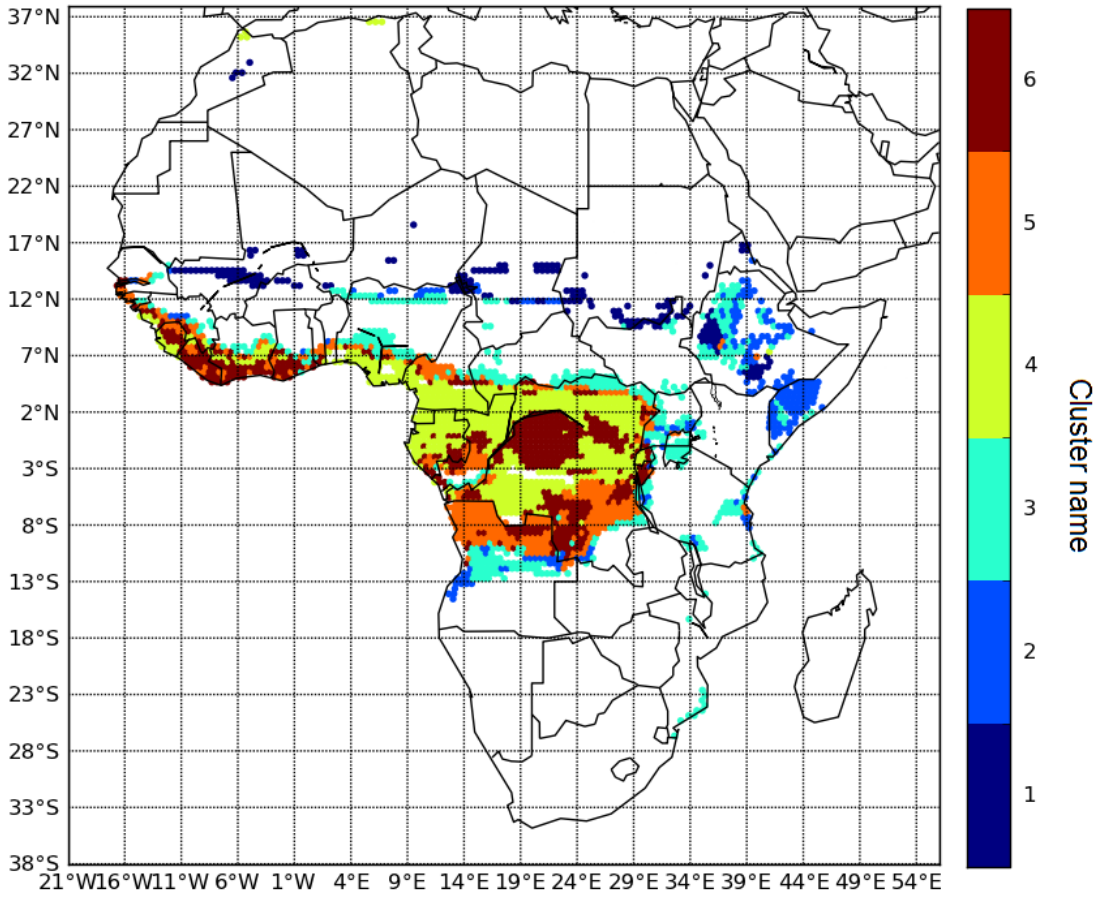

Figure 40: This figure shows cluster 2 with 6 sub-clusters on the map.

have the best prediction result on layer 1, 2 and 10 and tuned MHCMs can be the optimal models on layer 5 and layer 11. As shown in figure 47, tuned GMs is the best models for layer 11 and layer 13 and tuned HCMs can be the optimal model on layer 5. tuned AHCM is the optimal models on the rest of layers.

Figure 48 and Figure 49 shows RMSE of different layers projected on the map. Figure 48 shows that RMSE of tuned AHCMs on the map is also equatorial symmetric. The abnomal layer appears in layer 9 and layer 14 in the north of Africa and RMSE on those layers are at least 420. Figure 49 shows RMSE of tuned AHCMs on west part of data points of layer 7 and east part of data points of layer 7 is symmetric as symmetric axis is layer 7.

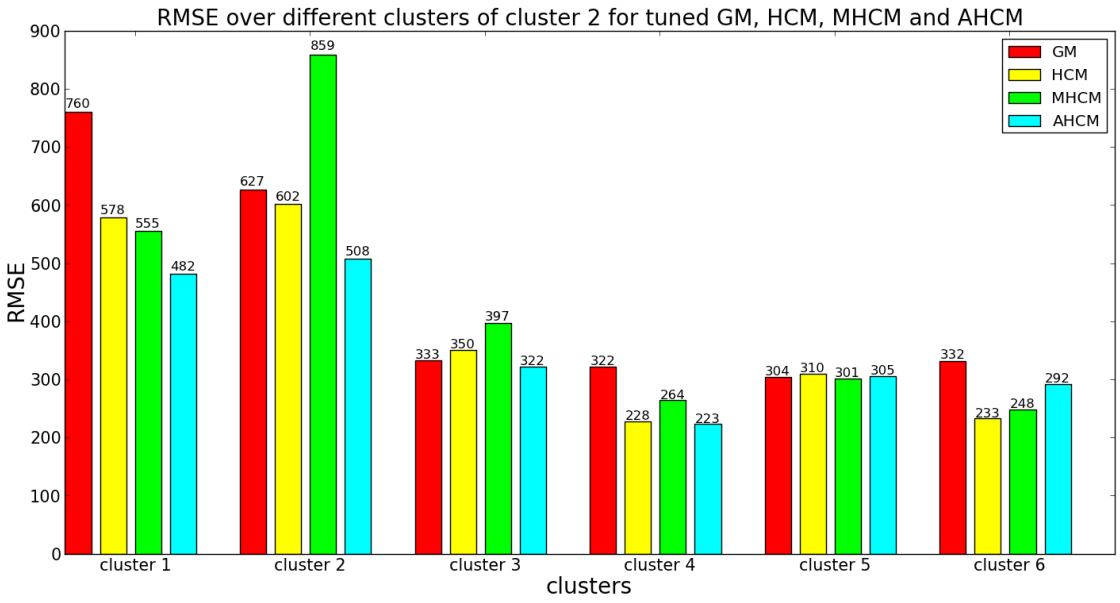

Figure 41: This figure shows RMSE of tuned GMs, HCMs, MHCMs and AHCMs for six sub-clusters of cluster 2 of Africa data

5.3 Discussion

This section is a short conclusion on comparison of global models and local models and the way to select different local models in different situations are also illustrated. Finally, it is a brief explanation of reasons why data points in the area around Turkana lake are difficult to make accurate predictions.

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 9
Subcluster 1	AHCM	AHCM			AHCM	GM		
Subcluster 2	HCM	AHCM			HCM	GM		
Subcluster 3	AHCM	AHCM			AHCM	HCM		
Subcluster 4	-	AHCM			AHCM	AHCM		
Subcluster 5	-	HCM	HCM	HCM	AHCM	-	HCM	GM
Subcluster 6	-	HCM			AHCM	-		
Subcluster 7	-	-			AHCM	-		
Subcluster 8	-	-			MHCM	-		
Subcluster 9	-	-			GM	-		
Subcluster 10	-	-			GM	-		

Table 22: This table shows optimal models on different clusters

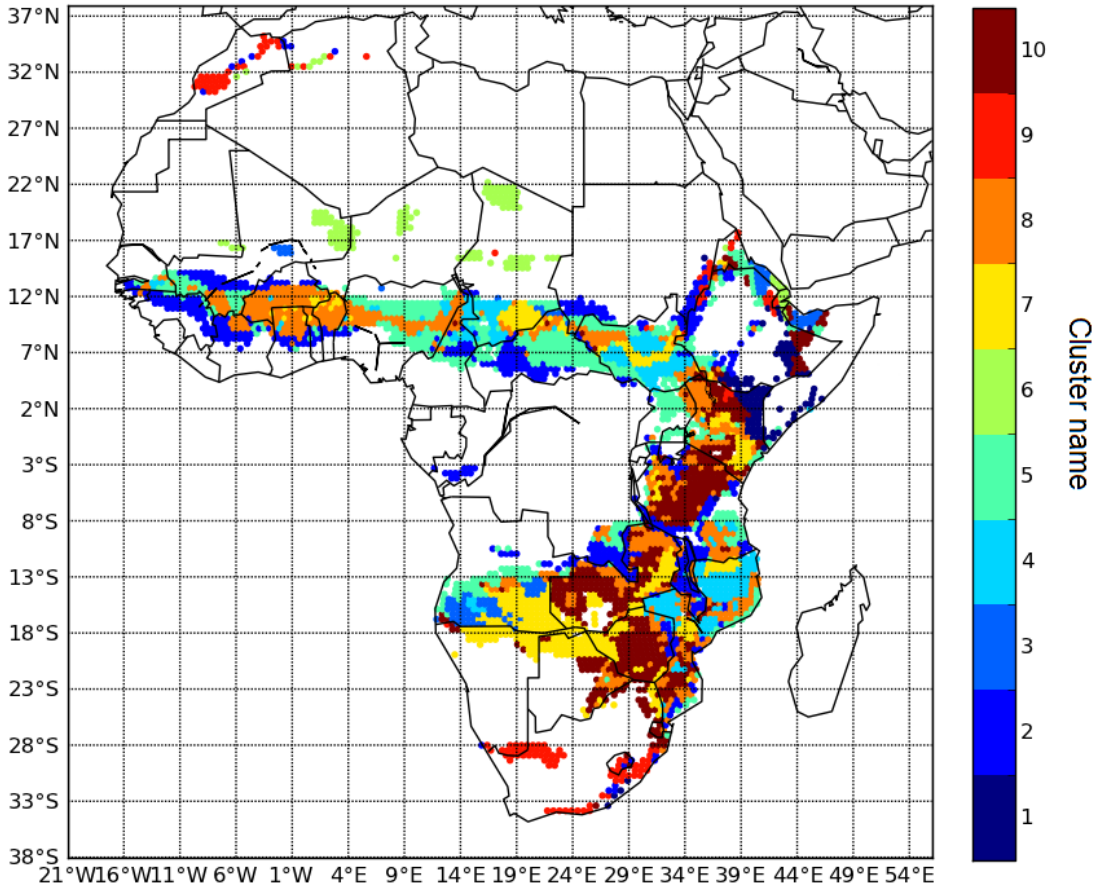

Figure 42: This figure shows cluster 5 with 10 sub-clusters on the map.

As shown in Table 23 and Table 24, rows in blue reveals performance of the baseline and rows in green are best models. Therefore, as for performance on the whole testing data, performance of AHCMs is the best. RMSE of AHCM after tuning parameters is 26.2% smaller than AHCM before tuning parameters. Meanwhile, AHCM before tuning parameters is the best among all models before tuning parameters and it is improved 7% compared with performance of global model before tuning parameters; Thus, this shows that clustering based local models can indeed improve performance of global models even though parameters are not tuned. Furthermore, performance of AHCM after tuning parameters is 31% better than performance of GM without tuning parameters.

Moreover, as mentioned in previous sections, although tuned AHCM has the best performance on the whole testing data in general, it is still possible that best tuned

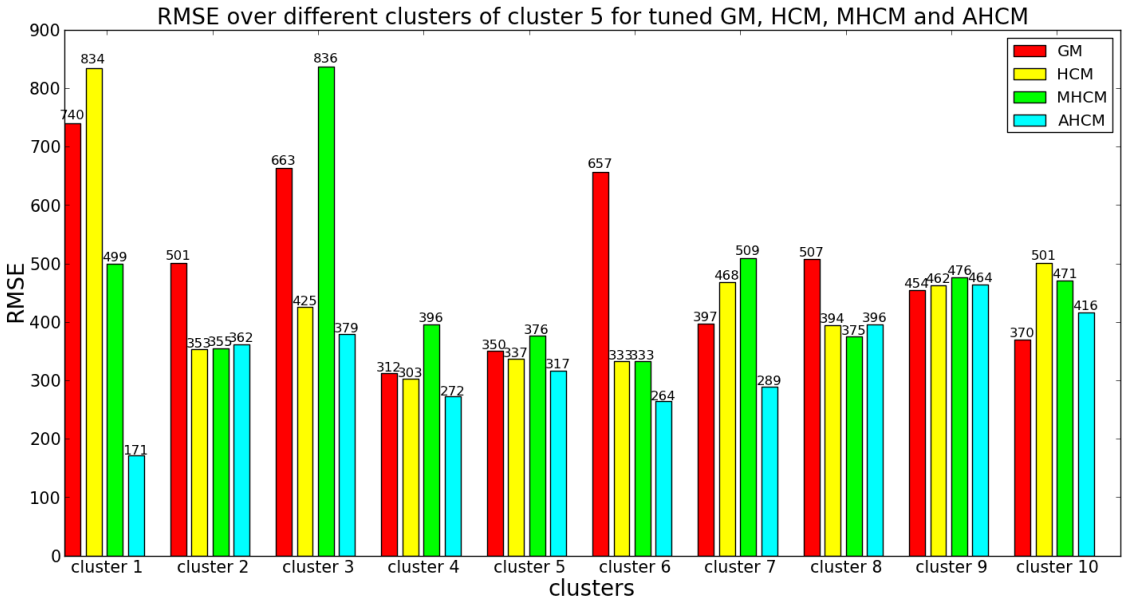

Figure 43: This figure shows RMSE of tuned GMs, HCMs, MHCMs and AHCMs for 10 sub-clusters of cluster 5 of Africa data

AHCM can have limitation in making predictions on some small parts of data. Table 22 shows optimal models on different clusters and it can be used as a reference for users to choose models based on different data points. For examples, if there is a group of new test data, those data points can be firstly clustered with all data points available to be 10 clusters. Assuming that all data points in that group are merged with sub-cluster 8 of cluster 5 in Africa, tuned MHCMs is the optimal models for that group of data points according the result in Table 22.

As shown in Table 22, tuned GM and MHCM is the best model instead of tuned AHCMs on sub-cluster 8, 9 and 10 of cluster 5 and area of Turkana basin consists of mainly subcluser 8, 9 and 10. Thus the best model for Turkana Basin is the combination of tuned GM and tuned MHCM. Moreover, RMSE of a small area around Turkana lake for 5 models are: GM: 526, HCM: 756, AHCM: 889, MHCM: 497, MHCM and GM: 489. RMSE of the combination of MHCM and GM is 55% of RMSE of AHCM and result of the best model on the area around Turkana lake is not a very accurate prediction. Figure 50 and Figure 53, shows prediction error on that small area of Turkana lake in Kenya. In those figures, RMSE are calculated in a small vertical slides with 3 data points so the whole testing data are partitioned into 2745 slides from the west to the east. Thus, each circle in those figures represents a

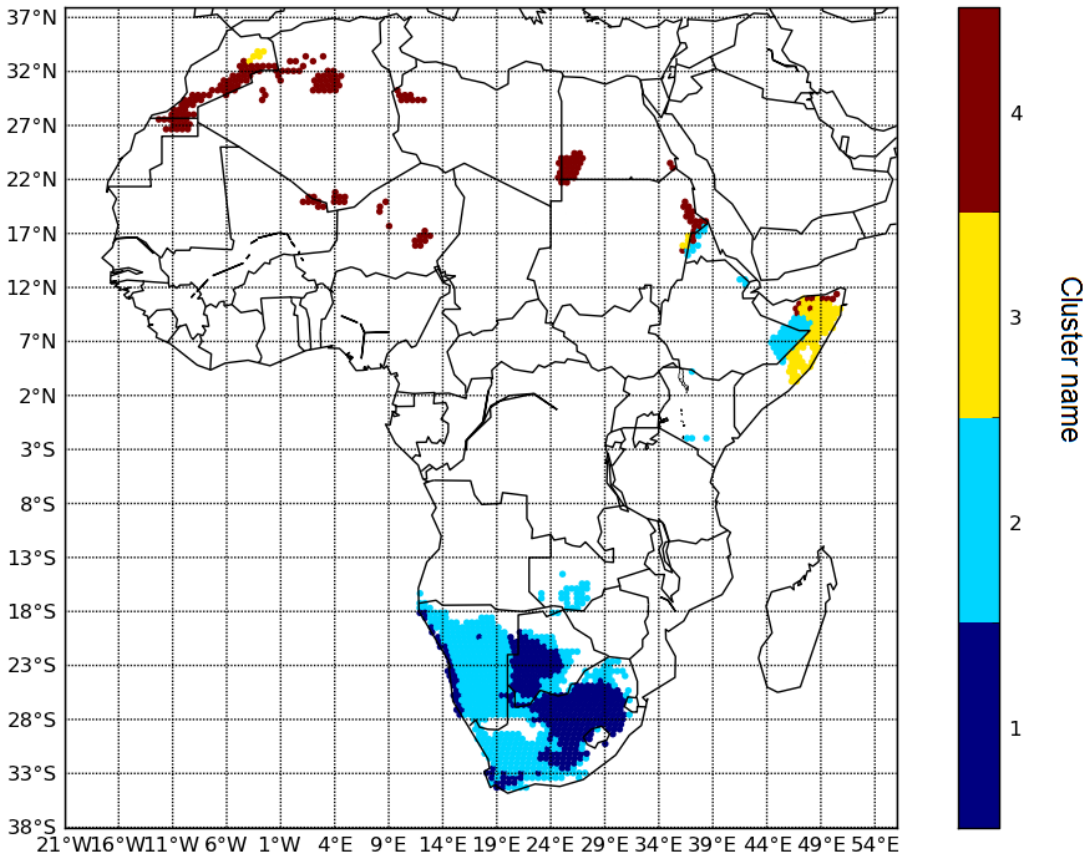

Figure 44: This figure shows cluster 6 with 4 sub-clusters on the map.

RMSE among three data points. The color map represent value of RMSE. For those 4 models, the trend is that RMSE around the Turkana lake increase large sharply and RMSE on data points that are not near the lake are not larger than 400. More importantly, prediction result of those models shows that NPP of Turkana basin is around 1000 but the real NPP of that area cannot be larger than 700. It is also rare that there are not much vegetation around a lake. It is possible that teeth features of plant-eating animals living in that area is very similar to plant-eating animals that lives in a humid area with NPP that is around 1000 because there is a lake in that area. Therefore, this can result in predictions that are below expectations.

Figure 55 shows the change of prediction error for tuned GM, HCM, MHCM and AHCM over the change of number of species of data points. The trend of RMSE of four models are almost the same. When number of species of a data point is equal to 5, performance of four models are the worst. Then RMSE of four models

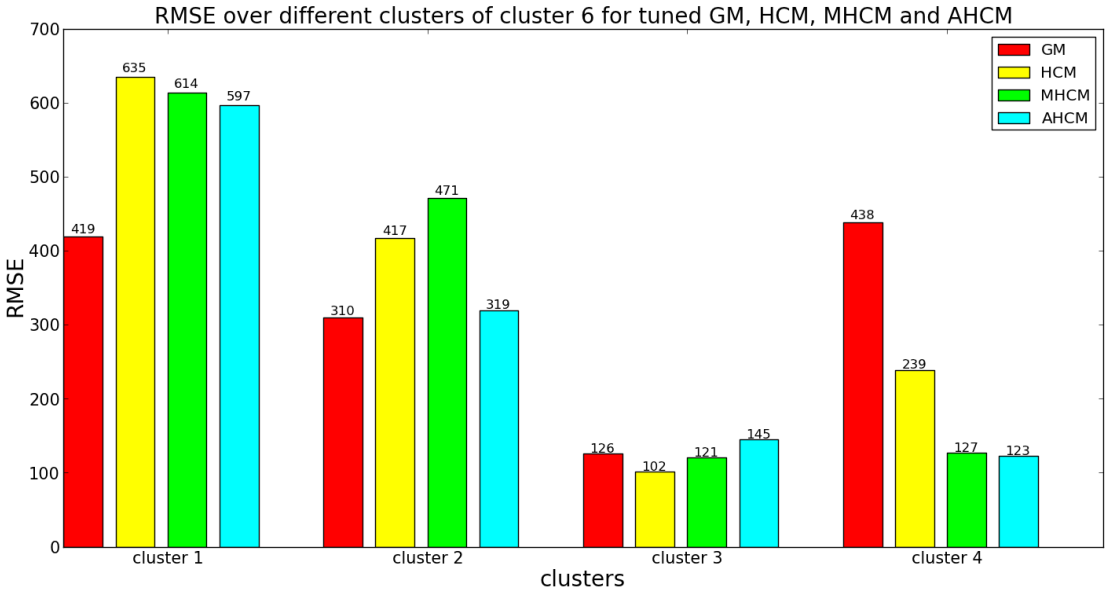

Figure 45: This figure shows RMSE of tuned GMs, HCMs, MHCMs and AHCMs for 4 sub-clusters of cluster 6 of Africa data

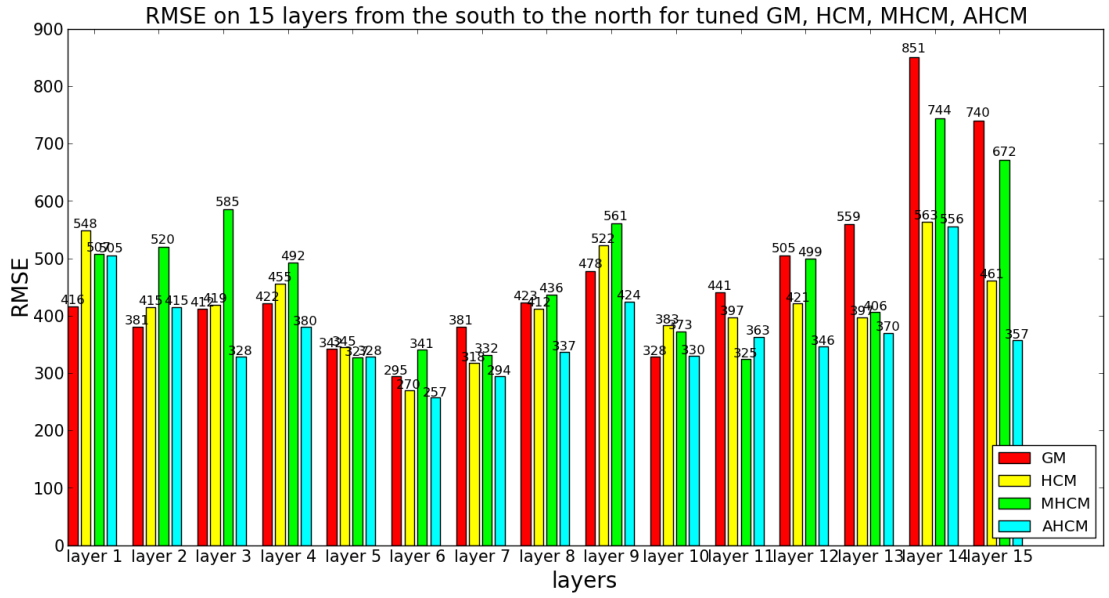

Figure 46: This figure shows performance of tuned GMs, HCMs, MHCMs and AHCMs on 15 horizontal layers with equal number of data points from south to north of Africa.

decrease with the change of number of species and when number of species is around 30, RMSE of GM, HCM and AHCM is the smallest. However, tuned MHCMs have

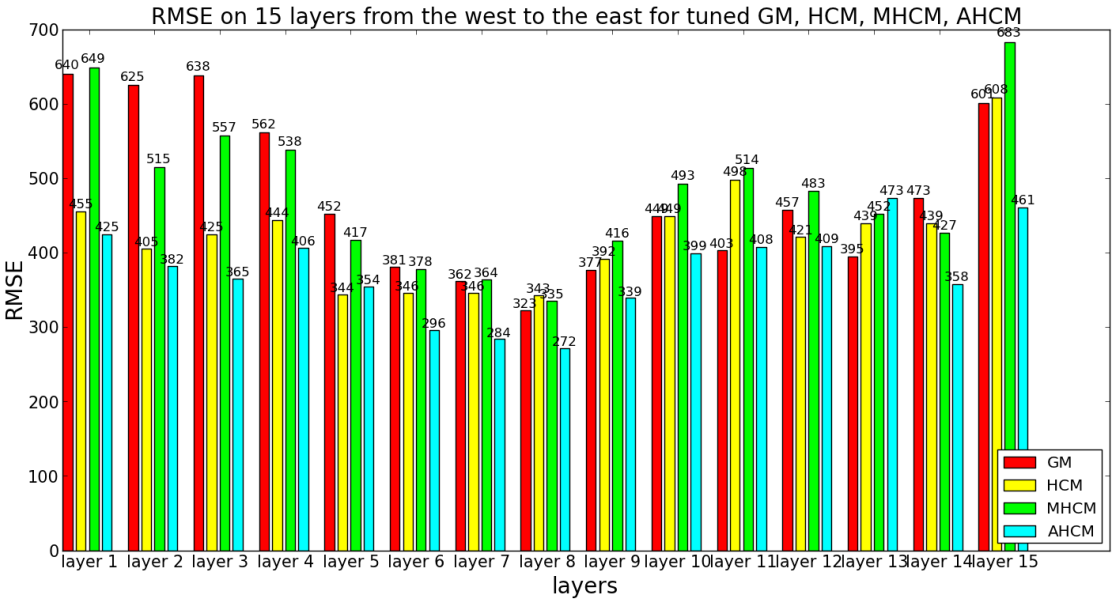

Figure 47: This figure shows performance of tuned GMs, HCMs, MHCMs and AHCMs on 15 vertical layers with equal number of data points from west to east of Africa.

the best performance when number of species is the largest. Figure 56 shows RMSE of tuned HCMs on the whole Africa. Thus, performance of models on the west side of Madagascar is much more worse than that on the east side of Madagascar. More importantly, as shown in Figure 54, number of species of the west side of Madagascar is around 5. Therefore, number of species on the west side of Madagascar is not sufficient so that performance of models on those data points are under expectations.

5.4 Evaluation procedures

Because of data is not i.i.d, r^2, RMSE and MAE are calculated over all data. Table 25 and Table 26 show results of all models in four different situations.

Moreover, compared the general performance of all models in standard 11 fold cross validation with performance of models in vertical spatial cross validation, models have more accurate prediction in standard 11 fold cross validation. This same trend also lies in comparison between standard leave-one-out cross validation and spatial leave-one-out cross validation. This trend appears because autocorrelated data of each group of testing data are pruned in vertical spatial cross validation and spatial

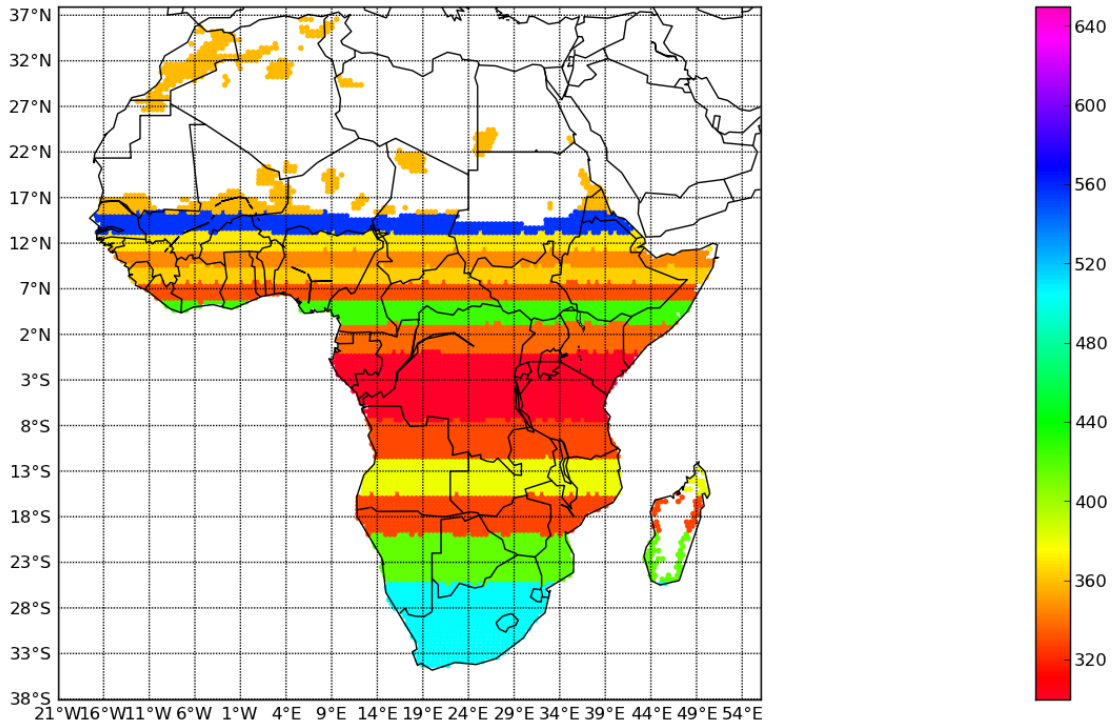

Figure 48: This figure shows RMSE of tuned AHCMs on each layer from the south to the north and map location of layers to Africa continent on the world map

leave-one-out cross validation. Furthermore, when compared a model in standard 11 folds cross validation with the same model in leave-one-out cross validation, the performance of that model is almost the same. However, when compared a model in vertical spatial cross validation with the same model in spatial leave-one-out cross validation, the performance of the model improves a lot in spatial leave-one-out cross validation. One of the reason can be that less data were discarded in spatial leave-one-out cross validation compared to number of data discarded in vertical spatial cross validation. However the running time of spatial leave-one-out cross validation is much more larger since the same number of models as the number of data are built in spatial leave-one-out cross validation. Rotation forest is not tested in leave-one-out cross validation since it took a long time. Therefore, pruning autocorrelated data reduces performance of models indeed. However, if the number of data is not that large, spatial leave-one-out cross validation can be a good choice.

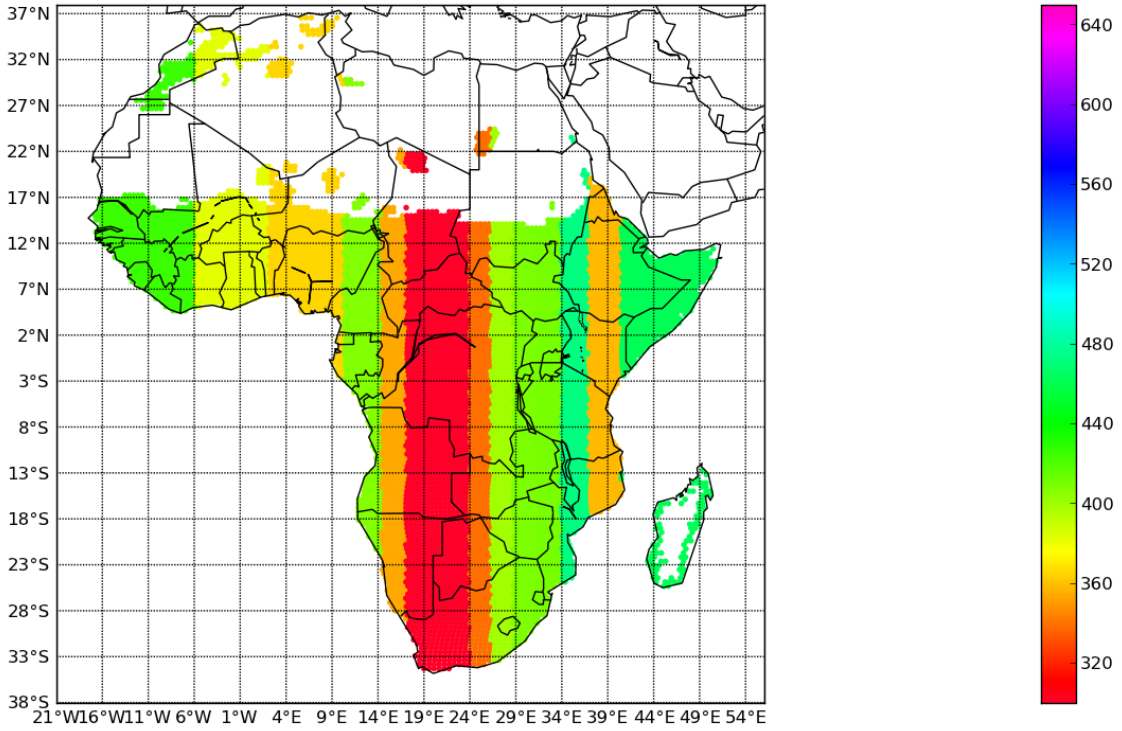

Figure 49: This figure shows RMSE of tuned AHCMs on each layer from the west to the east and map location of layers to Africa continent on the world map

6 Case study

According to experiment results described in previous sections, models are selected based on table 22. Thus, in the first step, we cluster the present day data and fossil data to be 10 clusters then we can select the optimal model. For fossil data, there are 5 clusters. Cluster 2, 4 and 5 are combined with the present day data and they are clustered again separately. Then we can clearly discover which cluster in fossil data are merged with which cluster in Africa data. Thus we can use the table to find the best model for making prediction on its corresponding cluster in fossil data. For example, a sub cluster of cluster 2 of fossil data can be merged with sub cluster 1 of cluster 5 in Africa data so the best model is AHCM. Then, the model with the same parameter settings as in the experiment for that cluster of Africa data are applied to the sub-cluster of cluster 2 of fossil data. This step is repeated until NPP of all fossil data are predicted.

model name	RMSE	MAE
baseline	565	430
GM	552	425
BM	737	572
MBM	557	412
HCM	550	422
MHCM	550	422
AHCM	515	392

Table 23: This table shows performance of best models without tuning parameters on the whole testing data.

model name	RMSE	MAE
baseline	565	430
GM	488	384
BM	652	495
MBM	533	402
HCM	429	335
MHCM	491	355
AHCM	380	291

Table 24: This table shows performance of tuned models on the whole testing data

Figure 58 to Figure 61 show prediction of NPP over time period. Figure 57 shows NPP in present data. In present day, the average NPP in Turkana Basin area is around 600 to 800. When time period starts from 0.01 to 2 Ma, the mean NPP is 1021. When time is from 2 to 3 Ma, the average NPP is 981. Thus the trend is when time changes from 0.01 to 3 Ma, the environment in Turkana Basin area becomes dry. However, when time is from 3 to 4 Ma, the mean NPP is 1123 and when time is from 4 to 7.8 Ma, the mean NPP is 1104. So from 3 Ma, the environment in Turkana Basin area starts become a little bit humid. Then from 4 to 7.8 Ma, the environment remains almost the same humidity. In addition, environment in time period of 3 Ma to 7.8 Ma is more humid than the environment in present day.

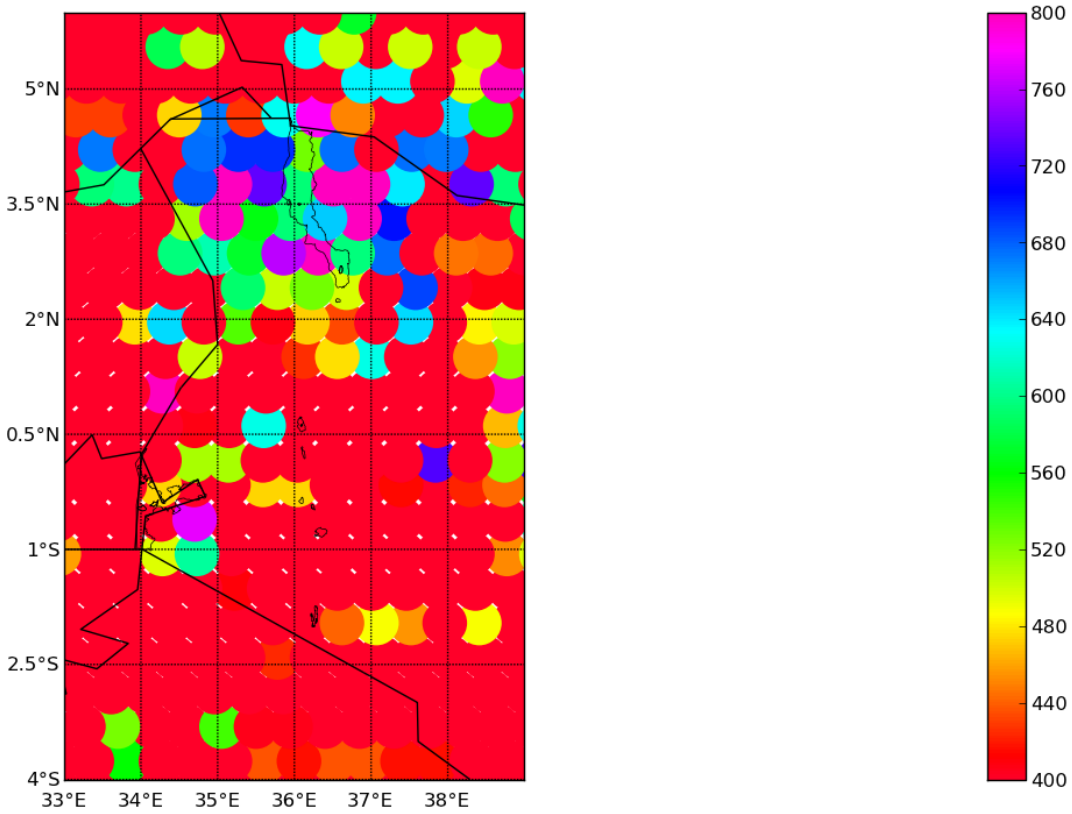

Figure 50: This figure shows RMSE of tuned AHCMs that are calculated in vertical slides with 3 number of data points from the west to the east of Africa and RMSE of data points around Turkana lake in Kenya are plotted

7 conclusions

This thesis is aimed to develop accurate models to build relationship between dental features of large plant-eating mammals and apply the model on fossil data to reconstruct the climate or environment in the ancient time between 0.01 and 7 million years ago. We can transfer models trained on dental features of mammals in present day to fossil data, because we use average traits of dental features on each site and fossil data share the same dental features with those of present data. Four datasets are utilized, animals occurrences data, animals dental features data, climate data and fossil data. By aggregating datasets of modern days, data point of a site has 8 mean dental features and terrestrial net primary productivity (NPP) as a climate variable. A mean dental feature of a site indicates the average trait

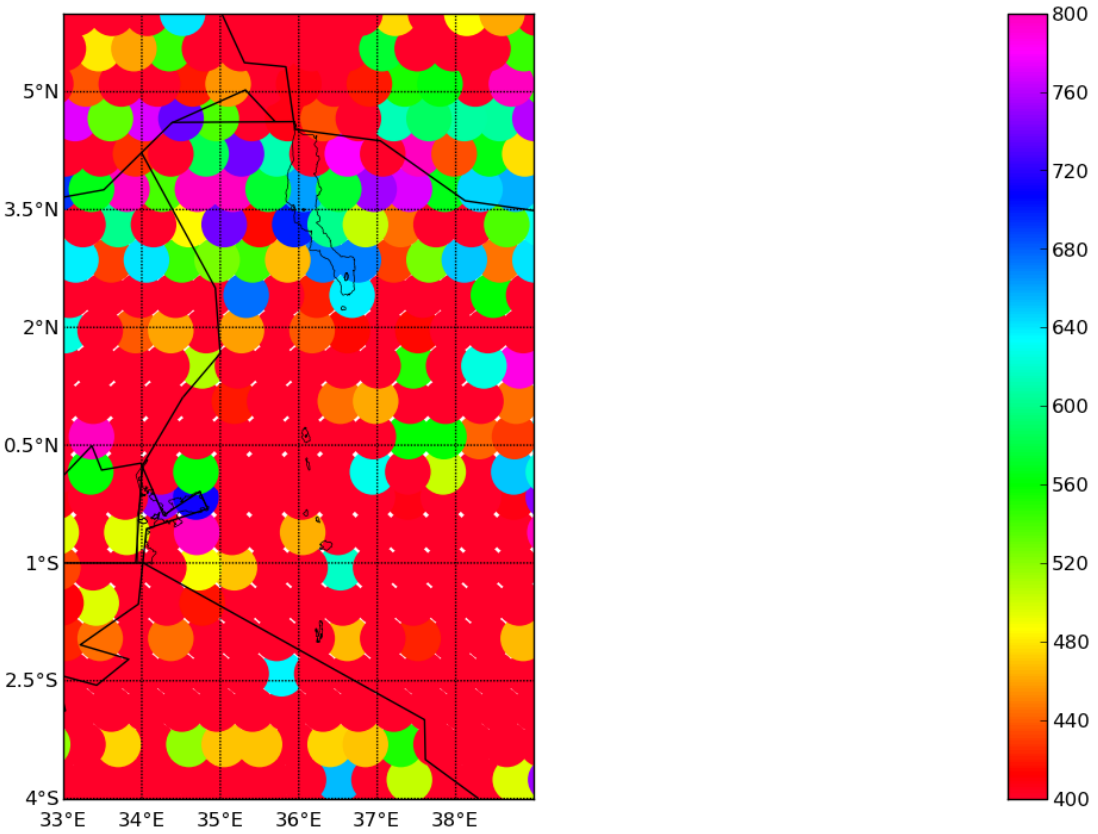

Figure 51: This figure shows RMSE calculated as the way described in figure 50 but they are RMSE of tuned HCMs

of a community of mammals that occurs on that site. Sites represent grid cells of 50km× 50km. The location of them are indicated by pairs of longitude and latitude.

In this thesis, data points with 8 dental features are input feature space and the response variable is NPP represents fixed energy stored in vegetation. Since input data points are not independently and identically distributed, generalization error may increase while amount of training data increases. Thus, we propose three types of local models: baseline models, hierarchical clustering based models and advanced hierarchical clustering based models. For baseline models, we propose two types of models: baseline models and modified baseline models. In baseline models, we select training data that has the same latitude as testing data. In modified baseline models, we select training data that has same latitude in both the Northern and Southern Hemisphere as testing data. In hierarchical clustering based models,

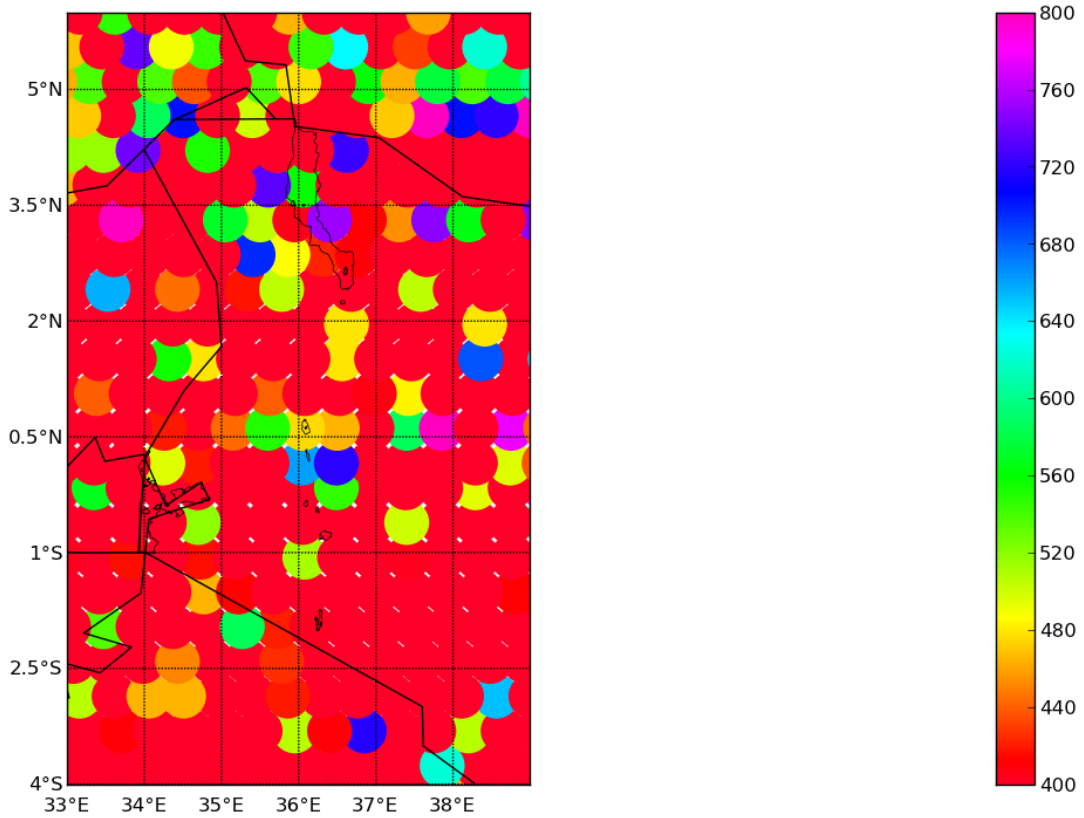

Figure 52: This figure shows RMSE calculated as the way described in figure 50 but they are RMSE of tuned GMs

we cluster training data and testing data by hierarchical clustering and we select k number of clusters in training data that match testing data closely. Furthermore, we also design a new strategy to optimize the prediction performance of hierarchical clustering based mode, a modified hierarchical clustering based model. We partition testing data into several layers with equal span in latitude and we apply optimal parameter settings for each layer. Ultimately, in advanced hierarchical clustering based model, we cluster testing data into several sub-clusters and the training data selection of each sub-cluster follows the same procedure in hierarchical clustering based models. Hierarchical clustering shows similarity between two clusters by Euclidean distance so that we can select number of clusters in training data in the order of similarity compared to testing data. The main idea of designing those local models is to select training data matching testing data the most.

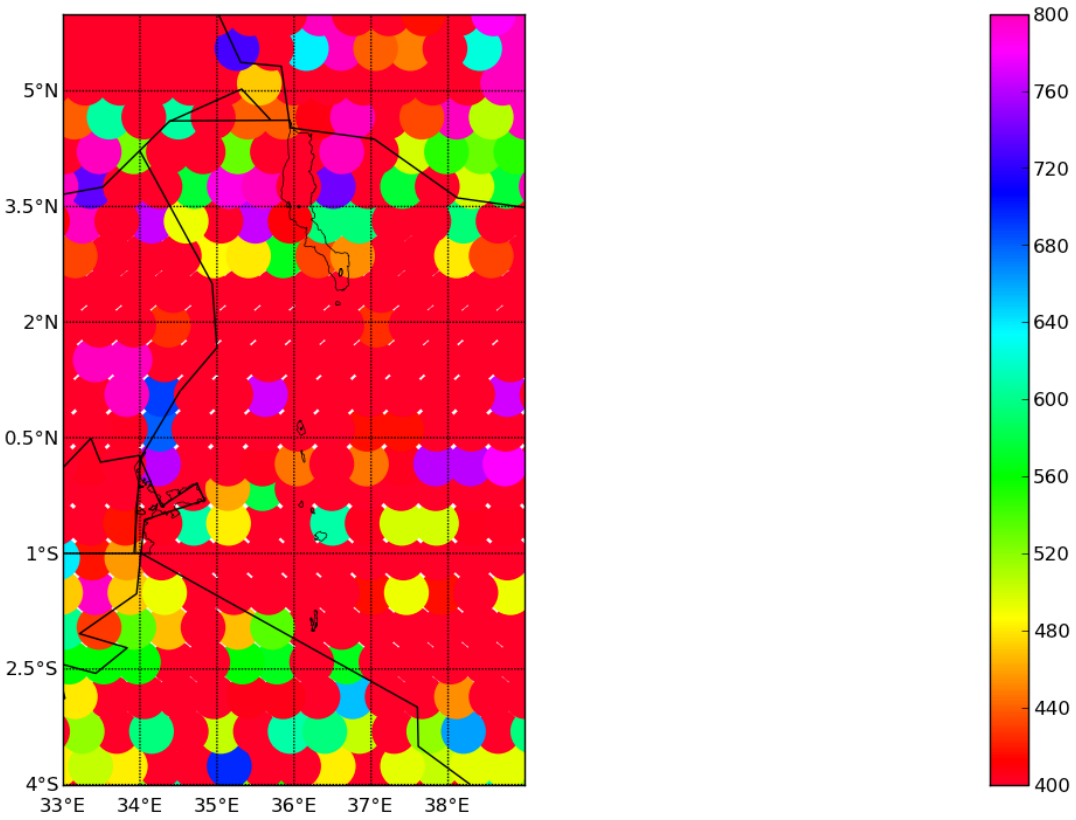

Figure 53: This figure shows RMSE calculated as the way described in figure 50 but they are RMSE of tuned MHCMs

Furthermore, we propose vertical spatial cross validation to solve spatial autocorrelation. In this thesis, standard cross validation can cause overfitting when tuning parameters of models. Thus, in vertical spatial cross validation, we partition data points to k test folds in the ascending order of longitude. Furthermore, data points at the distance less than 500km to either boundary of test folds are discarded. The remaining data points are used for training data or validation data. Finally, we measure performance of models with RMSE and MAE on unseen testing data.

In our experiments, global models are also tested for making comparison of our proposed local models. In addition, OLS, DT, RaF, GBR and RoF are utilised for building models and their performance are compared as well. Moreover, we select Africa continent as test continent since its climate is relatively least influenced by human activities and non-Africa data points forms training data pool. In addition,

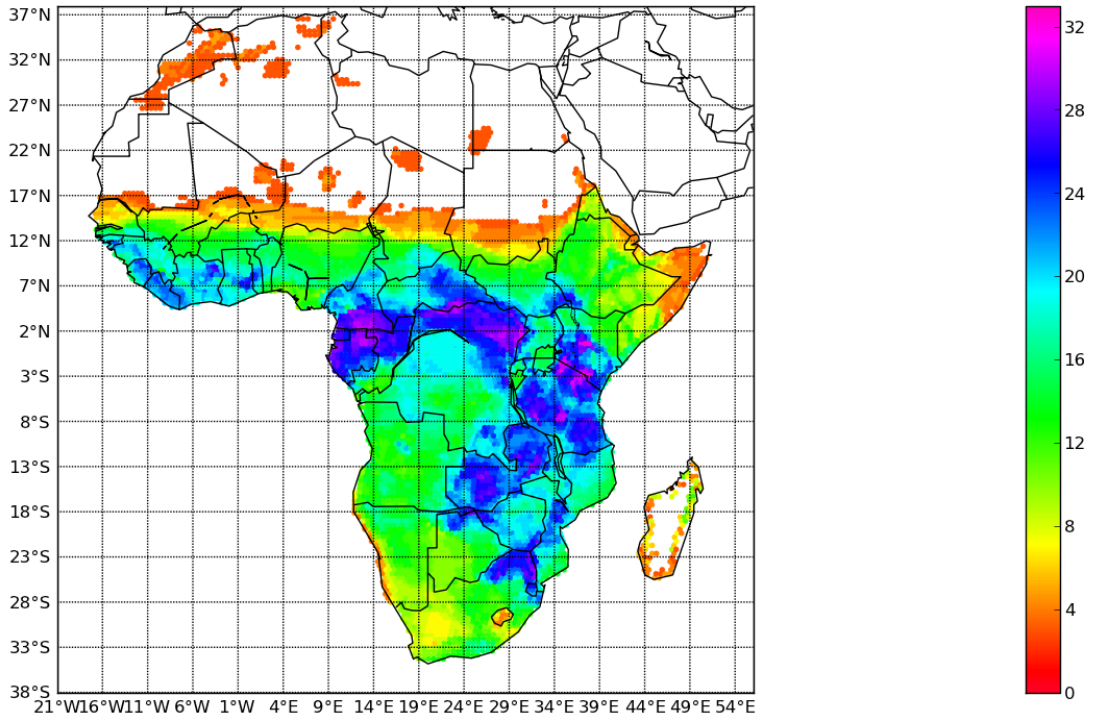

Figure 54: This figure shows distribution of species on the Africa continent.

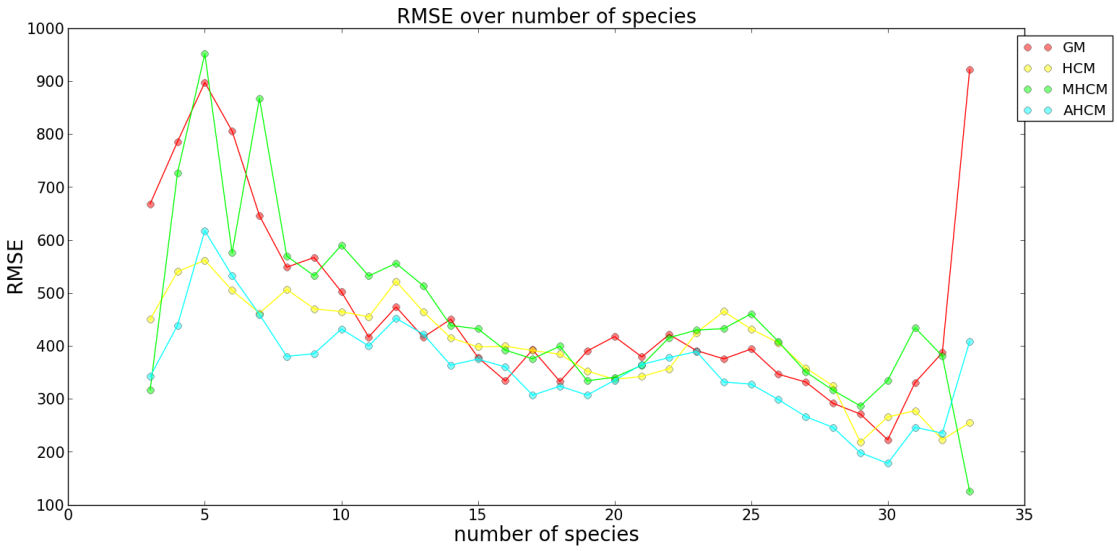

Figure 55: This figure shows RMSE of four tuned models over the change of number of species.

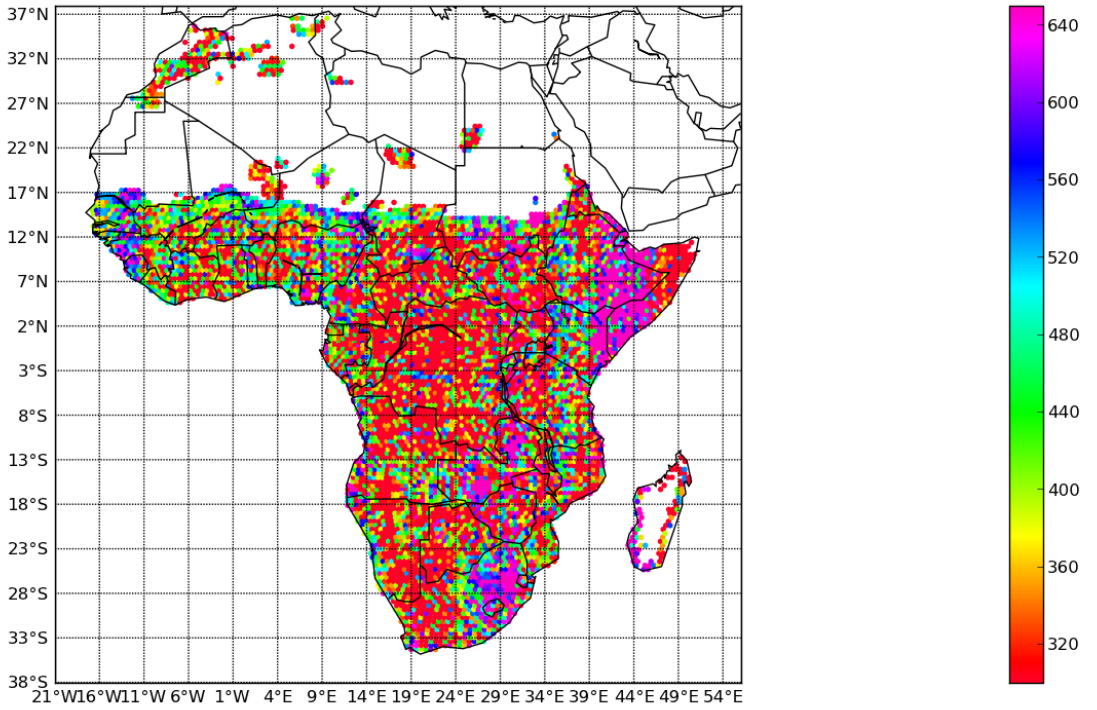

Figure 56: This figure shows RMSE of tuned HCMs on each layer with 3 data points from the west to the east and map location of layers to Africa continent on the world map

we choose the global model built by OLS as the baseline due to its simplicity and widely-application in related research work.

Furthermore, we conduct two experiments. In the first experiment, we build global models and local models using default parameter settings. For three clustering based models, we record all prediction results in the process of appending number of clusters in training data. In the second experiment, we tune parameters of local and global models using vertical spatial cross validation. Specifically, testing data are partitioned into 3 folds and the optimal parameters are obtained by minimising RMSE of models on validation data.

In the first experiment when parameters are not tuned, RMSE of the baseline is 565. The optimal model is AHCM with RaF and number of clusters in training data is 9 while the RMSE is 515. Therefore, this experiment suggests that our cluster-

standard 11-fold cross validation			
	r^2	RMSE	MAE
OLS	0.618	490	391
decision tree	0.863	294	198
rotation forest	0.740	405	326
gradient Boosting regressor	0.854	304	206
random forest	0.871	285	193

vertical spatial cross validation			
	r^2	RMSE	MAE
OLS	0.547	535	428
decision tree	0.639	477	352
rotation forest	0.649	471	387
gradient Boosting regressor	0.574	518	372
random forest	0.702	434	321

Table 25: result of 11 folds cross validation and vertical spatial cross validation

standard leave-one-out cross validation			
	r^2	RMSE	MAE
OLS	0.618	490	391
decision tree	0.863	294	197
gradient Boosting regressor	0.857	301	205
random forest	0.872	284	192

spatial leave-one-out cross validation			
	r^2	RMSE	MAE
OLS	0.587	511	408
decision tree	0.723	418	307
gradient Boosting regressor	0.701	434	315
random forest	0.764	385	286

Table 26: result of leave-one-out cross validation and spatial leave-one-out cross validation

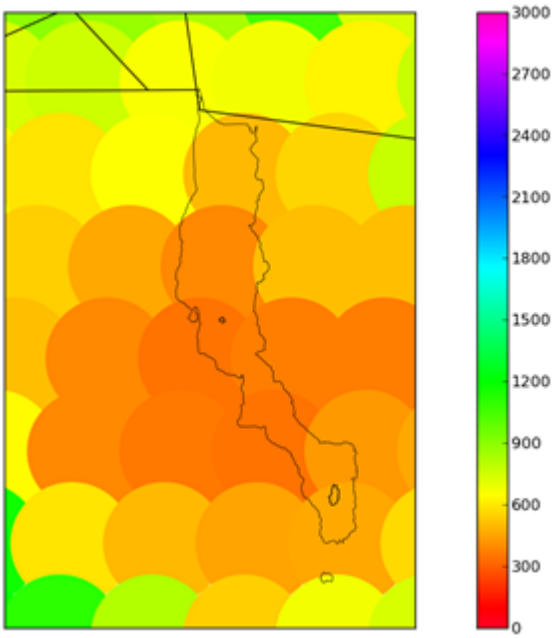

Figure 57: NPP at present day

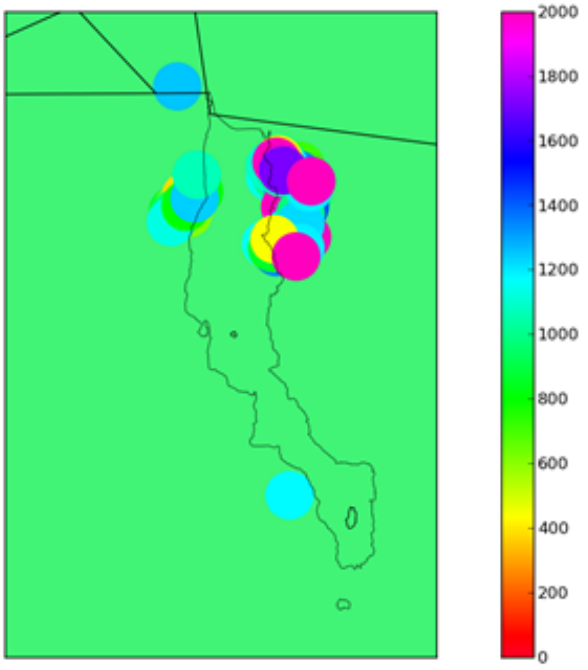

Figure 58: NPP from 0.01 to 2 Ma

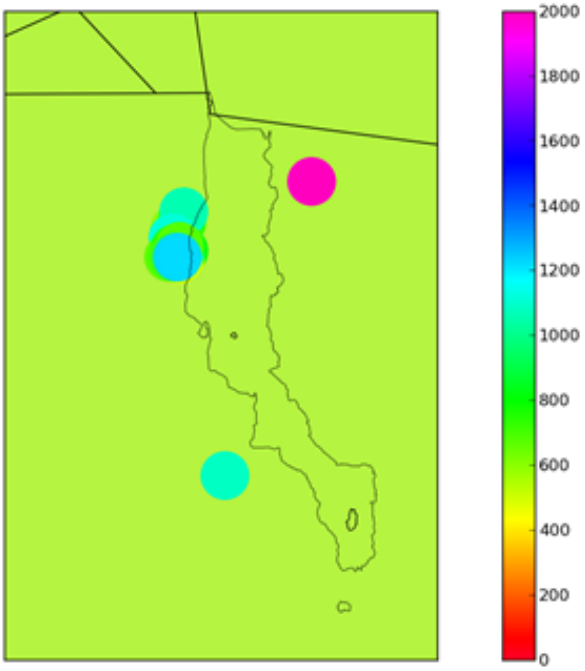

Figure 59: NPP from 2 to 3 Ma

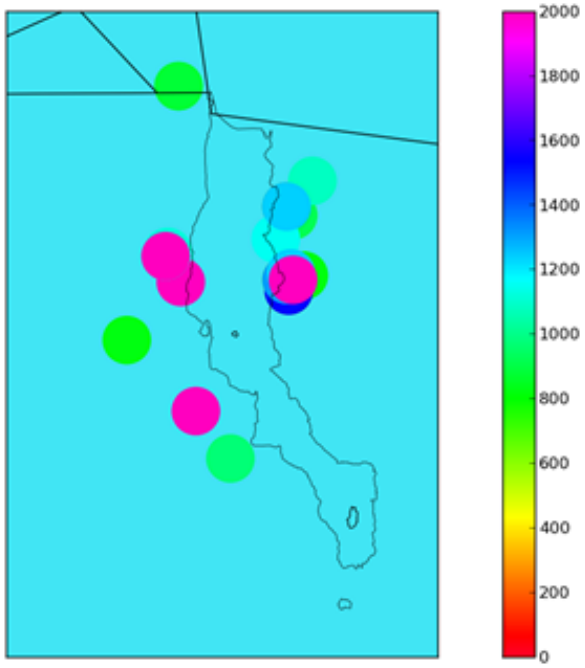

Figure 60: NPP from 3 to 4 Ma

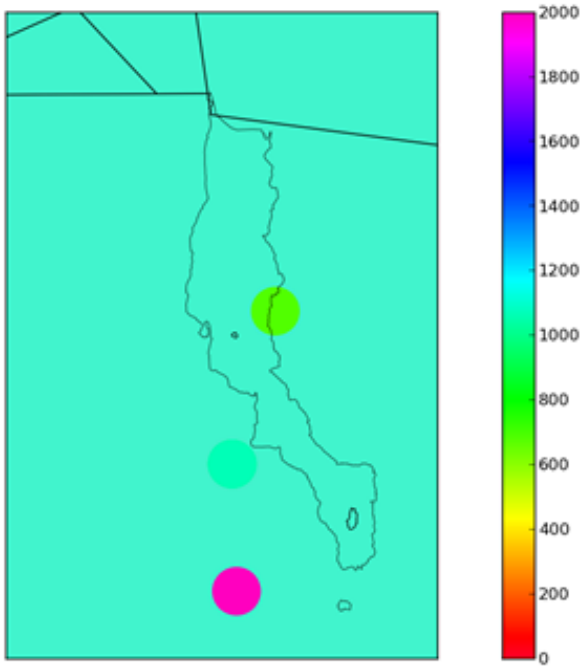

Figure 61: NPP from 4 to 7.8 Ma

ing based models can improve predication performance compared to global models. Since parameters of machine learning algorithms are not tuned, the improvement is only contributed by selecting training data that is similar to testing data.

In the second experiment, performance of the baseline is the same since the global model with OLS has no parameters. Considering RMSE on the whole testing continent, AHCM is the optimal model with RMSE 380 and it is reduced 33% compared to the baseline. However, if performance of models on different geological regions are compared, AHCM is not the optimal models on all clusters. Thus, the results suggest that there are no perfect models that can perform well in all geological regions. A scheme shows optimal models of different clusters in Table 22. This scheme can be utilized as following. For additional new testing data, we can append and cluster with the whole data points. For example, if the new testing data are all merged in sub-cluster 9 of cluster 5, tuned global models are the optimal models according to the cell on the 9th row and 5th column in the table.

Furthermore, if the testing continent are partitioned to 15 horizontal layers with equal number of data points, the RMSE of each layer is equatorial symmetric. In addition, for all models, layers near equator have lowest prediction error. Thus, this fact reveal that the prediction on data points in equatorial climate zone is most reliable. In two thirds of layers, AHCM has the best performance. The reason is that firstly, input features of testing data in a small cluster varies less. Moreover, in the algorithm, to select training data, we cluster the small testing cluster with data points in training data pool again. Thus this clustering result can be more reliable. Secondly, in the algorithm, we tune parameters and we can choose the optimal algorithms with the best parameters for each sub-cluster.

More importantly, GM is the optimal model for desert climate zone in the south of Africa. In addition, for the desert climate zone in the north, AHCM is the best model. One possible reason is that the input data space of mean dental traits and NPP on desert area in the south varies more which results in unstable area clustering result of AHCM. In this case, global models can achieve better results since there are more training data involved. Clustering-based models can achieve better results when variation of data points is small. Therefore, in this case, we can further cluster that area to obtain sub-clusters and improve clustering based models.

Moreover, we compare performance of global models and local models on region in Lake Turkana. It contains three sub-clusters of cluster 5: sub-cluster 8, 9 and 10. Thus, a combination of global model and MHCM can have the best performance with RMSE 497. Climate in Lake Turkana is different from the around area. Although there is a lake, its NPP shows that vegetation in that area is like desert. However, the existence of lake can influence dental features of nearby plant-eating mammals so dental features in that area is similar to relatively humid places. In AHCM, data points in Lake Turkana are clustered with data points in the relatively humid area. Therefore, AHCM have worse performance on them. Rare climate condition in Lake Turkana region can be a significant reason that data points in that area are difficult to predict.

In addition, we discover that data points in Madagascar are also difficult to predict. It is suspected that number of species of data points can also influence performance of models. For all models, when number of species on a site is around 4 or 5, all models have large prediction error that is at least 550. This conforms to the fact that performance of models on west Madagascar is much worse than the east side and number of species on the west side are around 4 or 5.

Furthermore, we compare vertical spatial cross validation that we proposed with spatial leave-one-out cross validation and standard cross validation. The results suggest that standard cross validation can cause overfitting in this setting indeed. When number of data points is very large, we recommend VSCV since the algorithm shortens the running time at the expense of larger RMSE. More precisely, RMSE of VSCV is 12.7% larger than SLOO.

Moreover, the result of case study shows that the trend of climate in Turkana Basin is that the environment firstly becomes dry slowly and it is the driest between around 2 to 3 Ma. Then, it starts becoming humid and tend to be stable. Climate between 4 Ma and 7 Ma is much more humid than than climate in present day in Turkana Basin.

Finally, we recommend two directions of future research. As considered in paper [H$^+$06], in our study, each prediction has the same cost. In order to avoid

predictions that have very large error compared to real value, different costs should be added in the process of model building. For example, if we define desert NPP as around 400, the NPP prediction of a data points in desert to be 100 or 500 does not change the fact that the environment is desert. But if the prediction of NPP is 1000, this prediction is meaningless since it shows that the environment of the data point is savannas or forest environment. Thus those types of prediction need to be avoided. Therefore, this can be a future research of our study. Moreover, in our work, all dental features are involved in building predictive models. The future work can design a new structural correspondence learning algorithm as in the paper [BMP06] and only use some pivot dental features to transfer models built on modern day data to fossil data.

References

ANC07 Arnold, A., Nallapati, R. and Cohen, W. W., A comparative study of methods for transductive transfer learning. *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*. IEEE, 2007, pages 77–82.

BCK+08 Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. and Wortman, J., Learning bounds for domain adaptation. pages 129–136.

BDP07 Blitzer, J., Dredze, M. and Pereira, F., Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. pages 440–447.

BHG+17 Barnosky, A. D., Hadly, E. A., Gonzalez, P., Head, J., Polly, P. D., Lawing, A. M., Eronen, J. T., Ackerly, D. D., Alex, K., Biber, E. et al., Merging paleobiology with conservation biology to guide the future of terrestrial ecosystems. *Science*, 355,6325(2017), page eaah4787.

BJOBK06 Bahn, V., J O'Connor, R. and B Krohn, W., Importance of spatial autocorrelation in modeling bird distributions at a continental scale. *Ecography*, 29,6(2006), pages 835–844.

BLY+10 Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J. and Elston,

D. A., Regression analysis of spatial data. *Ecology letters*, 13,2(2010), pages 246–264.

BMP06 Blitzer, J., McDonald, R. and Pereira, F., Domain adaptation with structural correspondence learning. *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2006, pages 120–128.

Bre12 Brenning, A., Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The r package sperrorest. *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*. IEEE, 2012, pages 5372–5375.

Dar09 Darwin, C., *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. John Murray, 2009.

DFBH03 Diniz-Filho, J. A. F., Bini, L. M. and Hawkins, B. A., Spatial autocorrelation and red herrings in geographical ecology. *Global ecology and Biogeography*, 12,1(2003), pages 53–64.

DI09 Daumé III, H., Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.

EPL+10a Eronen, J., Puolamäki, K., Liu, L., Lintulaakso, K., Damuth, J., Janis, C. and Fortelius, M., Precipitation and large herbivorous mammals i: estimates from present-day communities. *Evolutionary Ecology Research*, 12,2(2010), pages 217–233.

EPL+10b Eronen, J., Puolamäki, K., Liu, L., Lintulaakso, K., Damuth, J., Janis, C. and Fortelius, M., Precipitation and large herbivorous mammals ii: application to fossil data. *Evolutionary Ecology Research*, 12,2(2010), pages 235–248.

FEJ+02 Fortelius, M., Eronen, J., Jernvall, J., Liu, L., Pushkina, D., Rinne, J., Tesakov, A., Vislobokova, I., Zhang, Z. and Zhou, L., Fossil mammals resolve regional patterns of eurasian climate change over 20 million years. *Evolutionary Ecology Research*, 4,7(2002), pages 1005–1016.

FEL+03 Fortelius, M., Eronen, J., Liu, L., Pushkina, D., Tesakov, A., Vislobokova, I. and Zhang, Z., Continental-scale hypsodonty patterns, cli-

matic paleobiogeography, and dispersal of eurasian neogene large mammal herbivores. *Deinsea*, 10,1(2003), pages 1–12.

Fri01 Friedman, J. H., Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

FŽK+16 Fortelius, M., Žliobaitė, I., Kaya, F., Bibi, F., Bobe, R., Leakey, L., Leakey, M., Patterson, D., Rannikko, J. and Werdelin, L., An ecometric analysis of the fossil mammal record of the turkana basin. *Phil. Trans. R. Soc. B*, 371,1698(2016), page 20150232.

GTFŽ17 Galbrun, E., Tang, H., Fortelius, M. and Žliobaitė, I., Computational biomes: the ecometrics of large mammal teeth.

H+06 Hand, D. J. et al., Classifier technology and the illusion of progress. *Statistical science*, 21,1(2006), pages 1–14.

HAB15 Hempson, G. P., Archibald, S. and Bond, W. J., A continent-wide assessment of the form and intensity of large mammal herbivory in africa. *Science*, 350,6264(2015), pages 1056–1061.

HCP+05 Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. and Jarvis, A., Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25,15(2005), pages 1965–1978.

HDFMB+07 Hawkins, B. A., Diniz-Filho, J. A. F., Mauricio Bini, L., De Marco, P. and Blackburn, T. M., Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. *Ecography*, 30,3(2007), pages 375–384.

HGB+07 Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B. and Smola, A. J., Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 2007, pages 601–608.

Hij12 Hijmans, R. J., Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, 93,3(2012), pages 679–688.

JHF96 Jernvall, J., Hunter, J. P. and Fortelius, M., Molar tooth diversity, disparity, and ecology in cenozoic ungulate radiations. *Science*, 274,5292(1996), pages 1489–1492.

JWHT14 James, G., Witten, D., Hastie, T. and Tibshirani, R., *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014.

LPE+12 Liu, L., Puolamäki, K., Eronen, J. T., Ataabadi, M. M., Hernesniemi, E. and Fortelius, M., Dental functional traits of mammals resolve productivity in terrestrial ecosystems past and present. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20120211.

LRPB13 Le Rest, K., Pinaud, D. and Bretagnolle, V., Accounting for spatial autocorrelation from model selection to statistical inference: Application to a national survey of a diurnal raptor. *Ecological Informatics*, 14, pages 17–24.

LRPM+14 Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J. and Bretagnolle, V., Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23,7(2014), pages 811–820.

Mec17 Mechenich, M., Best practices for ecometric analysis: a case study correlating climate conditions and herbivore teeth in africa.

ODW+01 Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C. et al., Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51,11(2001), pages 933–938.

PKY08 Pan, S. J., Kwok, J. T. and Yang, Q., Transfer learning via dimensionality reduction. *AAAI*, volume 8, 2008, pages 677–682.

PPNH17 Pohjankukka, J., Pahikkala, T., Nevalainen, P. and Heikkonen, J., Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31,10(2017), pages 2001–2019.

PY10 Pan, S. J. and Yang, Q., A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22,10(2010), pages 1345–1359.

RK Ruß, G. and Kruse, R., Regression models for spatial data: An example from precision agriculture. *Advances in Data Mining. Applications and Theoretical Aspects*, 28, pages 450–463.

RKA Rodriguez, J. J., Kuncheva, L. I. and Alonso, C. J., Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28,10, pages 1619–1630.

Val84 Valiant, L. G., A theory of the learnable. *Communications of the ACM*, 27,11(1984), pages 1134–1142.

Ž16 Žliobaitė, Indrė and Rinne, Janne and Tóth, Anikó B and Mechenich, Michael and Liu, Liping and Behrensmeyer, Anna K and Fortelius, Mikael, Herbivore teeth predict climatic limits in kenyan ecosystems. *Proceedings of the National Academy of Sciences*, page 201609409.

Zli16 Zliobaite, Indre and Tatti, Nikolaj, A note on adjusting R^2 for using with cross-validation. *arXiv preprint arXiv:1605.01703*.

ŽPEF17 Žliobaitė, I., Puolamäki, K., Eronen, J. T. and Fortelius, M., A survey of computational methods for fossil data analysis. *Evolutionary Ecology Research*, 18,5(2017), pages 477–502.

ZW10 Zhang, H. and Wang, Y., Kriging and cross-validation for massive spatial data. *Environmetrics*, 21,3-4(2010), pages 290–304.